Molecular Genetics of Addiction and Related Heritable Phenotypes

Genome-Wide Association Approaches Identify "Connectivity Constellation" and Drug Target Genes with Pleiotropic Effects

George R. Uhl,[a] Tomas Drgon,[a] Catherine Johnson,[a] Chuan-Yun Li,[a,b] Carlo Contoreggi,[c] Judith Hess,[c] Daniel Naiman,[d] and Qing-Rong Liu[a]

[a]*Molecular Neurobiology Branch, National Institutes of Health (NIH), Intramural Research Program (IRP), National Institute on Drug Abuse (NIDA), Baltimore, Maryland, USA*

[b]*Center for Bioinformatics, College of Life Sciences, Peking University, Beijing, China*

[c]*Office of the Clinical Director, NIH-IRP (NIDA), Baltimore, Maryland, USA*

[d]*Department of Mathematics, Johns Hopkins University, Baltimore, Maryland, USA*

Genome-wide association (GWA) can elucidate molecular genetic bases for human individual differences in complex phenotypes that include vulnerability to addiction. Here, we review (a) evidence that supports polygenic models with (at least) modest heterogeneity for the genetic architectures of addiction and several related phenotypes; (b) technical and ethical aspects of importance for understanding GWA data, including genotyping in individual samples versus DNA pools, analytic approaches, power estimation, and ethical issues in genotyping individuals with illegal behaviors; (c) the samples and the data that shape our current understanding of the molecular genetics of individual differences in vulnerability to substance dependence and related phenotypes; (d) overlaps between GWA data sets for dependence on different substances; and (e) overlaps between GWA data for addictions versus other heritable, brain-based phenotypes that include bipolar disorder, cognitive ability, frontal lobe brain volume, the ability to successfully quit smoking, neuroticism, and Alzheimer's disease. These convergent results identify potential targets for drugs that might modify addictions and play roles in these other phenotypes. They add to evidence that individual differences in the quality and quantity of brain connections make pleiotropic contributions to individual differences in vulnerability to addictions and to related brain disorders and phenotypes. A "connectivity constellation" of brain phenotypes and disorders appears to receive substantial pathogenic contributions from individual differences in a constellation of genes whose variants provide individual differences in the specification of brain connectivities during development and in adulthood. Heritable brain differences that underlie addiction vulnerability thus lie squarely in the midst of the repertoire of heritable brain differences that underlie vulnerability to other common brain disorders and phenotypes.

Key words: **pleiotropic; cell adhesion; Monte Carlo**

Address for correspondence: George R. Uhl, Molecular Neurobiology, Box 5180, Baltimore, MD 21224. Voice: +410-550-2843x146, fax: +410-550-1535. guhl@intra.nida.nih.gov

Ann. N.Y. Acad. Sci. 1141: 318–381 (2008). © 2008 New York Academy of Sciences.
doi: 10.1196/annals.1441.018

Introduction

Genome-wide association (GWA) is now increasingly the method of choice for identifying allelic variants that contribute to complex genetic disorders, especially those with *polygenic* bases (i.e., those derived from effects at many gene loci, each with modest effects, as well as from environmental determinants; see also Glossary).[1–9] Substance dependence was one of the first complex phenotypes for which replicated association-based genome scanning data were reported.[3,10–14] A torrent of information is now available from GWA studies of a number of other complex, brain-based phenotypes that display substantial heritability and are unlikely (based on linkage study results) to result from many common gene variants that produce large effects.[1,2,4,5] A number of these other heritable, brain-based phenotypes co-occur with addictions and are thus good candidates to display genetic overlaps with addiction.

No single approach to designing GWA studies or to analyzing GWA data is now universally accepted. There is now no universal standard for considering GWA results significant in ways that allow us to identify polygenic allelic variants in reasonably sized single experiments. Here, we describe specific sets of working hypotheses about the genetic architecture of addiction (e.g., the vulnerability to develop dependence on an addictive substance). This set of hypotheses is also useful for considering the molecular genetic bases for other common, complex phenotypes that, like addictions, display both substantial evidence for heritability and little evidence for large influences from any single gene (e.g., single-gene, Mendelian influences or oligogenic effects that come from a few genetic loci, each with moderate effects on the phenotype). We then detail experimental design and analytic approaches that arise from working hypotheses about underlying genetic architecture and likely sources of false positive results.

A number of samples provide the bases for these analyses. We focus on clusters of genomic markers whose allele frequencies distinguish control individuals from those with substance dependence or addiction-related phenotypes. We describe identification of chromosomal regions that contain clusters of such nominally positive results in replicate samples for addiction vulnerability. We then describe evidence for generalization that arises from identification of overlapping chromosomal locations of clustered positive results for different phenotypes. These data thus support *pleiotropic* influences (i.e., contributions of the same allelic variants to multiple phenotypes) of common allelic variants on several of the brain-based phenotypes. The data thus document overlapping heritable influences on several interesting brain phenotypes.

We focus here on clinical phenotypes that co-occur with addiction and a structural brain phenotype: individual differences in frontal cortical volume. Twin studies document sizable heritable components for individual differences in the volumes of brain regions. High heritabilities are especially evident for individual differences in frontal and temporal cerebral cortical regions.[15] Volumes of these brain regions have been reported to be reduced in substance-dependent individuals.[16–19] Increasing evidence from functional magnetic resonance imaging (fMRI) and positron emission tomography (PET) studies of individuals with substance dependence and related phenotypes identifies functional differences in these brain regions.[20,21] We thus focus on this "frontal cortical volume" phenotype.

A number of the genes identified here encode classical "druggable" targets for pharmacological modulation, including enzymes, receptors, and transporters. Other genes encode cell adhesion–related molecules. We discuss genes in each of these classes below.

Utility of GWA in Examining the Molecular Bases of Heritable Influences on Vulnerability to Addiction and Related Phenotypes

One way to view GWA is in relationship to linkage-based genome scanning because most

of the efforts to positionally clone gene variants for complex human disorders (e.g., those that are likely to be caused by multiple genetic and environmental factors) have used linkage-based methods. Linkage asks how addiction phenotypes and genetic markers (typically genotyped approximately every 1/400th–1/1,000th of the genome) move together through pedigrees of closely related individuals. Chromosomal regions that contain marker alleles that move through pedigrees together with the trait are said to be linked to the trait. Many loci with nominally significant linkage to addiction phenotypes have been identified[22–31] (further references in[32]). Several of the loci identified in independent linkage studies do overlap. However, the large numbers of reported linkage-based studies of addictions yield large numbers of nominally positive results that cover virtually all chromosomes. These widespread results may not converge more than expected by chance, as we have documented in recent analyses of the reported data for linkage to smoking.[32]

Such inconsistent linkage data are consistent with the idea that most of the genetic architecture for human addiction vulnerability is polygenic in most populations. A growing consensus now holds that GWA (also termed *whole-genome association* or *association genome scanning*) is more likely than linkage approaches to yield positive results in polygenic complex disorders, such as addictions.[1,3,10–12,33–37] Association asks how addiction phenotypes and genetic markers (genotyped approximately every 1/500,000th to every 1/1,000,000th of the genome in current data sets) are found together in nominally unrelated individuals (although we are all distantly related to each other, of course). We and others have developed these methods, relying on the increasing densities of single nucleotide polymorphism (SNP) markers that can be assessed using SNP chip microarrays of increasing sophistication.[10–12,33,34,36,37] GWA gains power as densities of genomic markers increase. Association identifies much smaller chromosomal regions than linkage-based approaches, thus allowing us to identify variants in specific genes rather than in large chromosomal regions. GWA fosters pooling strategies that preserve confidentiality and reduce costs, as we discuss below.[10–12,33,38–44] GWA also provides ample genomic controls. Proper genomic controls can minimize the chances that disease versus control differences are confounded by occult stratification, such as the stratification that might arise from unintended occult ethnic mismatches between disease and control samples.

A Probable Underlying Genetic Architecture for Addiction and Other Heritable Brain-based Phenotypes

We approach analyses of the molecular genetic bases of addiction and related disorders from perspectives that are based on sets of working hypotheses concerning the underlying genetic architectures of these disorders and phenotypes. In general, the best experimental design and analytic approaches will probably arise from the best possible working hypotheses concerning (a) the genetic architectures of the disorders or phenotypes being evaluated, (b) the population genetics of the samples being tested, and (c) the anticipated association signals. It is also desirable to consider and provide controls for alternative hypotheses that might explain systematic differences between disease and control samples. Alternative hypotheses include (a) unintended stratification, such as that based on racial or ethnic differences between disease and control samples, as noted above; (b) uneven distribution of noise in some assays such that SNPs with the largest variance might be identified rather than the SNPs whose allelic frequencies truly differ between disease and control individuals; (c) stochastic, chance differences between disease and control samples (at least some of these are highly likely in any single study that uses many repeated comparisons); and (d) sampling issues such that genetics related to the ways in which the samples are ascertained and obtained (e.g., features such as differential willingness to consent) are identified

rather than true disease versus control differences. Many of these concerns become more prominent as hundreds of thousands or millions of repeated comparisons are made using single sample sets. Moreover, many of these concerns may become more acute as attempts to rapidly assemble larger and larger sample sizes drive investigators to include subsamples that may well contribute occult heterogeneities to overall disease and/or control samples (see below).

Family, Adoption, and Twin Data that Support Substantial Polygenic Heritability for Addictions

Current models for the genetic architecture of substance dependence in the population are based on information from (a) family, adoption, and twin data that support substantial heritability for addictions; (b) twin data, in which concordance in genetically identical monozygotic and genetically half-identical dizygotic twins are compared that document that most of this heritable influence is not substance-specific; and (c) linkage-based and GWA studies that fail to provide evidence for genes of major effect (i.e., for any single gene whose variants produce substantial differences in addiction vulnerability) for substance dependence.

Support for the idea that vulnerability to addictions is a complex trait with strong genetic influences that are largely shared by abusers of different legal and illegal addictive substances[45–48] comes from classical genetic studies. Family studies document that first-degree relatives (e.g., siblings) of addicts display a greater risk for developing substance dependence than more distant relatives.[45,49] Adoption studies find greater similarities in levels of substance abuse between adoptees and biological relatives than between adoptees and members of the adoptive families.[45] In twin studies, differences in concordance between genetically identical and fraternal twins also support heritability for vulnerability to addictions.[47,50–56] Twin data allow quantitation of the amount of addiction vulnerability, about half, that is heritable. Twin data also support the idea that the environmental influences on addiction vulnerability that are not shared among members of twin pairs are much larger than those that are shared by members of twin pairs (e.g., $e^2 \gg c^2$ in virtually every such study). Most environmental influences on human addiction vulnerability are thus likely to come from outside of the immediate family environment (Fig. 1).

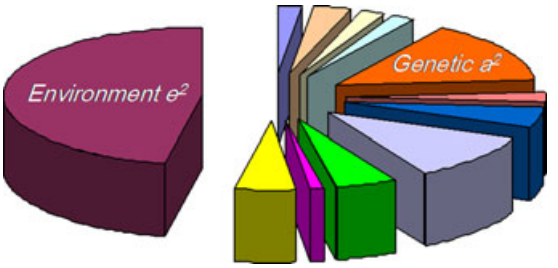

Figure 1. Pie graph model for the genetic architecture of human vulnerability to dependence on addictive substances. Polygenic additive genetic influences and environmental influences that are largely those that are not shared between members of sibships are depicted. Potential roles for g × g and g × e interactions are not depicted here.

Twin Data Document that Most of this Heritable Influence Is Not Substance-Specific, but Provides Higher-Order Pharmacogenomics

We are fortunate to have data from studies of identical versus fraternal twin pairs that evaluate the degree to which one twin's dependence on a substance enhances the chance that his or her co-twin will become dependent on a substance of a different class. Results of these analyses document that most of the genetic influences on addiction vulnerability are common to dependence on multiple different substances, although others do appear to be substance-specific.[46,53,54]

Elsewhere[32] we have suggested the following levels of analysis for pharmacogenomics and pharmacogenetics: (a) *primary pharmacogenomics*, which describes the genetics of individual differences in the adsorption, distribution, metabolism, and/or excretion of a drug; (b)

secondary pharmacogenomics, which describes individual differences in drug targets, such as the G protein–coupled receptors, transporters, and ligand-gated ion channels that are the primary targets of opiates, psychostimulants, and barbiturates, respectively; and (c) *higher-order pharmacogenomics*, which addresses individual differences in postreceptor drug responses. Such postreceptor drug responses are more likely to be common to actions of abused substances that come from several different chemical classes and act at distinct primary receptor or transporter sites in the brain. Based on the twin data currently available, we thus postulate that much, if not most, of the human genetics of addition vulnerability represents higher-order pharmacogenomics.

Failure to Document Evidence for Substance-Dependence Genes of Major Effect in Most Populations

Few careful studies have examined the ways in which most human addiction vulnerabilities move through families (e.g., *segregation analyses*). No such study indicates a major gene effect on addiction vulnerability in most current populations, with one exception: the "flushing syndrome" variants at the aldehyde (ALDH) and alcohol dehydrogenase loci in Asian individuals do provide genes of major effect in this population. Individuals with these gene variants are at lower risk for becoming dependent on alcohol than individuals with other genotypes[57] in Chinese,[58,59] Korean,[60] Japanese,[61–66] and other populations.[67,68] Homozygous ALDH2∗2 individuals are strongly protected from alcohol dependence.[61,62] This locus thus provides a good example of primary pharmacogenomics, though in a restricted population.

Quantity–frequency data for smoking also provide evidence for a replicable secondary pharmacogenomic effect of moderate magnitude. Several studies have shown that markers in the chromosome 15 gene cluster, which encodes the α3, α5, and β4 nicotinic acetylcholine receptors, display different allelic frequencies in heavy versus light smokers.[3,35,69] This chromosome 15 locus probably provides a good example of secondary pharmacogenomics because it has not been associated as reproducibly with dependence on other substances.

Linkage-based analyses for addiction vulnerabilities would be expected to reproducibly identify many of the genes whose variants exert major influences on human addiction vulnerability. However, existing linkage data for human dependence on alcohol, nicotine, and a number of other substances fail to provide any highly reproducible results that would support a role for any major gene locus ([22–31] and references in[32]). These results add to the conclusion that no locus individually appears to contribute a large fraction of the vulnerability to dependence on any addictive substance, with one caveat: these data come from subjects with largely European ethnic or racial backgrounds.[22,23,70–84] Nevertheless, as with many complex human disorders in which initial hopes for a tractable (e.g., *oligogenic*) underlying genetic architecture supported the use of linkage approaches, the linkage peaks that are identified in each individual study may be more likely to arise on other bases when the underlying architecture is in fact polygenic. Apparent linkage signals identified in single studies might result from polygenic influences from several genes that happen to lie near each other on human chromosomes or are found on stochastic bases when there is no true major effect from any single gene variant, for example.[85]

Current Models for the Genetic Architecture of Human Dependence

Current models for the genetic architecture of human dependence on legal and illegal addictive substances in the population thus postulate that each is affected roughly 50% by polygenic influences, that is, by variants in more than one individual gene, each of which contributes modest amounts to this overall genetic vulnerability. Such genetic architectural models posit that many of these genetic vulnerabilities

increase the risk for addiction to several pharmacologic classes of abused substances, but that some of these genetic influences are specific to drugs of one class.[32]

Analyses of twin data for vulnerability to develop dependence on a substance fit with large, additive genetic components (a^2), large components for nonshared environmental influences (e^2), and small components for c^2 terms that represent familial or other environmental influences that are shared between members of the twin pair.[47,50–56] What about the possibility that large interactions could occur between these genetic and environmental terms (G × E interactions), invalidating additive models for genetic and environmental contributions? G × E correlations of three types have been described.[86,87] In one terminology, *passive* G × E correlation occurs when parents transmit both genes and environmental influences that are relevant for a trait,[88,89] *active* G × E correlation occurs where subjects of a certain genotype actively select environments that are correlated with that genotype, and *reactive* G × E correlation occurs when an individual's genotype provides different reactions in response to the environment. Small values for c^2 influences of common environments shared by members of sibpairs appear to provide evidence against passive G × E correlations. On these bases, active and reactive G × E correlations remain of theoretical interest. However, one influential notion[88,90] suggests that G x E correlations are best regarded as parts of the genetic variance because "the non-random aspects of the environment are. . . consequence(s) of the genotype(es)."

Large interactions between genetic and environmental components would probably lead to (a) differences in estimates of heritability from samples obtained in different environments and (b) differences in molecular genetic findings in individuals from different environments. Data from studies of twins who were sampled from a number of different environments is nevertheless largely convergent. Such convergence supports relatively modest upper limits on (G × E) interactions between genetic and environmental influences on addiction vulnerability. Modest G × E influences are also consistent with molecular genetic results that identify substantial overlaps between the molecular genetics of vulnerability to dependence on illegal substances in samples from substantially different environments, such as the United States and Asia (see below).

Gene–gene interactions (G × G) of some magnitude appear likely, *a priori*, to make at least some contributions to addiction vulnerability. However, in the presence of substantial *epistasis*, G × G interactions in which specific alleles at one gene locus are required for expression of the effects of allelic variants at a second gene locus, segregation analysis data might provide uneven patterns of familiality. With considerable epistasis, second-degree relatives (e.g., cousins) of addicts would be less likely to display specific combinations of G × G alleles than first-degree relatives (e.g., siblings). Substance-dependence rates would thus drop more precipitously between first- and second-degree relatives of addicts than they would if most risk alleles exerted largely independent effects on addiction vulnerability.

Only limited family data allow for a comparison of concordance in first- versus second-degree relatives. However, the existing evidence does not support less concordance in second-degree relatives than we would anticipate based on the observed concordance in first-degree relatives and the assumption that most risk alleles produce largely independent effects.[91]

The Genetic Architecture for Substance Dependence in Individuals

What about the genetic architecture for substance dependence in individuals? Both *between-locus heterogeneity* and *within-locus heterogeneity* are likely. If we follow the implications of polygenic models for addiction vulnerability, we can infer that each dependent individual might display a nearly distinct set of risk-elevating or risk-reducing allelic variants. As an illustrative

example, we might postulate that (a) an individual must display at least 50 risk alleles to robustly elevate his or her likelihood of acquiring a substance-dependence disorder and (b) 200 genes contain common allelic variants that can augment addiction risk. Under such circumstances, it is easy to see that the exact genetic recipe for addiction vulnerability found in one addicted individual might be replicated in only a relatively small number of other addicted individuals. Such an underlying genetic architecture would be consistent with the failure of linkage-based methods to provide reproducible results in addictions because linkage relies on the identification of consistent patterns in the ways that specific DNA markers and phenotypes move through many families that display high densities of the disorder.

As noted above, the best-documented genetic heterogeneity for addictions comes from the chromosome 4 major gene effects found in poorly alcohol-metabolizing ("flushing") Asian individuals.[61–63,67] The best-documented substance-specific influence comes from the chromosome 15 nicotinic acetylcholinergic receptor gene cluster. Other examples of between-locus genetic heterogeneity and of genes whose variants exert substance-specific effects on use and/or dependence probably exist but have yet to be elucidated.

We also postulate that within-locus heterogeneity is likely, though not yet clearly documented in addiction, to our knowledge. Many common Mendelian disorders and rarer Mendelian phenocopies of common disorders display substantial heterogeneity within their pathogenic loci. For example, a number of variants of the same CFTR gene produce cystic fibrosis disorders,[92] and α synuclein missense variants and copy number variants can each provide phenocopies of idiopathic Parkinson's disease.[93] Evidence for within-locus heterogeneity in complex disorders is just beginning to accrue; such evidence now includes data from neurexin gene family variants in autism.[94–97]

Epigenetics and Individual Differences in Vulnerability to Addiction and Related Phenotypes

Epigenetics is now used with both classical and recently revised definitions. Classical definitions of *epigenetic* emphasize influences of variations that are not encoded in the primary DNA sequence but are nevertheless inherited; in other words, "a change in the state of expression of a gene that does not involve a mutation, but that is nevertheless inherited in the absence of the signal (or event) that initiated the change."[98] However, more recent definitions of *epigenetic* emphasize gene regulatory mechanisms that do not alter the primary DNA sequence and deemphasize the documentation of heritability.[98]

In the context of this review, heritable epigenetic influences are most relevant. One example of a classical, heritable epigenetic influence is imprinting, in which information is conveyed from parent to child through mechanisms that include DNA methylation or histone acetylation. These mechanisms retain the primary DNA sequence but can dramatically alter the function of specific genes. DNA methylation at CpG sequences in the promoter regions of genes can profoundly alter gene transcription. Because methylation during the course of maternal oocyte (or paternal sperm) development is key to this process, familial patterns of gender-specific transmission can provide evidence for this subset of heritable epigenetic influence. The modest quality of current family data sets for addiction renders them a relatively weak basis for any strong inferences concerning parent-of-origin effects. Nevertheless, no segregation data of which we are aware supports strong parent-of-origin effects on substance dependence. Thus, although there are obvious and large roles for nonheritable epigenetic influences in the biology of addiction, no current compelling evidence exists for any strong effects of overall heritable epigenetic influences, as classically defined. We nevertheless need to

be alert for such influences as we unravel the effects of variants in specific genes.

The Nature of the Allelic Variants Likely to Contribute to Individual Differences in Vulnerability to Addiction and Related Phenotypes

The analytic strategies described below are based on postulates that *common disease* or *common allele* models hold for many of the variants that alter vulnerabilities to addiction and related phenotypes. Rare variants may also explain significant fractions of the genetic risk for common diseases. However, increasing evidence supports roles in addiction vulnerability for allelic variants that are currently common and that are thus likely to be old in an evolutionary sense. Data indicating that such variants can be identified in diseased individuals from European, African, and Asian genetic backgrounds also point, in general, to variants of substantial age. How could *genetic selection* act on such common functional allelic variants over the large number of generations that are implied by this substantial age? It is conceivable that some currently common allelic variants could exert polygenic influences on addiction vulnerability without exerting positive or negative selective effects during lengthy evolutionary histories. However, it also seems likely that many allelic variants that influence addiction vulnerability can provide *balancing selection*; in other words, the effects of these variants may be favorable in some individuals, organs, or circumstances and unfavorable in others. Balancing selection might thus maintain relatively high frequencies of multiple functional allelic variants in the population over long periods of time.

The biology of some genes might allow for common, functional allelic variants that could escape selective pressures or exert balancing selection over many generations. However, other genes might not be able to harbor such allelic variations without engendering selective pressures that would reduce the frequency of one of the allelic variants in the population over time. Common allelic variants that are able to influence addiction vulnerability are thus likely to be restricted to a subset of genes whose products are involved in addictive processes. An important consequence of this logic follows: if a gene fails to display variants that influence vulnerability to addiction, the gene's products are not at all excluded from involvement in addiction. On the other hand, convincing data implicating a gene's common variants in addiction should prompt us to consider mechanisms whereby such variants might provide balancing (i.e., both positive and negative) selective influences in the differing environments through which the ancestors of current human populations have passed.

How does this discussion of common disease or common allele hypotheses relate to the postulates of genetic heterogeneity noted above? None of the points in the above discussion about common alleles and common variants precludes, or even reduces the likelihood of, contributions of rarer (or even private) allelic variants, including those that have arisen more recently in evolutionary time. Recently arising variations would be much more likely to persist for a number of generations in the face of even moderately negative influences on survival or fertility. Indeed, based on experience with other genetic disorders, it may be worthwhile to actively search for effects of rarer *phenocopy* variants in genes that are initially identified based on common, and evolutionarily older, allelic variants.[99] A rarer copy number variant might contribute to addiction vulnerability by altering the levels of expression of a gene that also has more common allelic variants that alter expression via SNPs in other gene elements, for example.[94–97] Such considerations support searches within identified loci for molecular genetic heterogeneity relevant to addiction.

In the analyses presented in this review, we focus on addiction-associated allelic variants that lie within genes. Evolutionarily old common haplotypes (i.e., groups of nearby variants that travel together through generations) that

lie within genes are among the most likely to be tagged by SNP markers that are represented on current microarrays. Haplotypes that involve genes are thus among the most likely variants to exist in currently reported data sets. It seems reasonable to postulate that many of these allelic variants that lie within genes provide regulatory variants that alter expression or regulation. Other variants are likely to alter mRNA halflives or mRNA splicing. Variants that alter mRNA splicing could occur at the locus of the affected gene (*cis*) or at genes at different loci that alter generic mRNA splicing processes (*trans*). Reproducible association of A2BP1 gene variants with addiction vulnerability, for example,[12] provide a good candidate for *trans* effects on mRNA splicing because this gene's product regulates splicing and thus is likely to modify the functions of a number of other genes expressed in the brain. It also seems likely that a minority of the addiction-associated variants will involve missense effects on expressed proteins.

It also seems likely that many addiction-associated variants will lie outside of genes, at least as we currently understand them. Loci reproducibly associated with diabetes and body mass, for example, lack the conventional hallmarks of genes, such as expressed sequences.[7] Although the analyses in this review focus on the identification of variants within genes, we should also remain alert for the roles of intergenic variations in chromosomal regions that lie between currently understood genes.

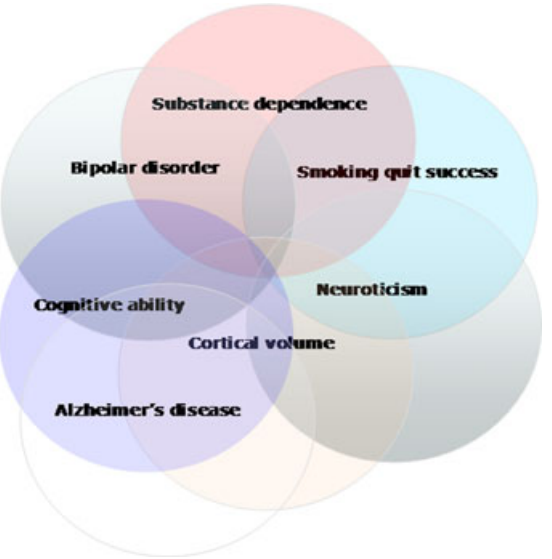

Figure 2. Venn diagram of overlapping genetic contributions to several of the phenotypes discussed here based on GWA data sets of about 500,000 SNPs. Note that the area of overlap in the figure does not necessarily represent the extent of overlap in the data sets. See text for more details.

Samples for Genome Studies of Human Addiction Vulnerabilities and Related Phenotypes

Sample 1: European American Polysubstance Abusers and Controls

European Americans volunteered to be subjects in research conducted at the National Institutes of Health (NIH) National Institute on Drug Abuse (NIDA) Intramural Research Program (IRP) in Baltimore, Maryland, based on word-of-mouth referrals and newspaper advertisements. Volunteers self-reported their ethnicities, provided drug-use histories, and provided *Diagnostic and Statistical Manual* (*DSM*) diagnoses of substance-use disorders.[10,100,101] "Abusers" displayed heavy lifetime use of illegal substances[33] and dependence on at least one illegal substance. "Controls" displayed neither abuse nor dependence on any addictive substance and reported no significant lifetime histories of use of any addictive substance. Individuals with intermediate levels of lifetime substance use without dependence were thus not included in analyses of substance dependence, although a number of them were included among samples studied for cognitive abilities (see Samples 15 and 16, below). Control individuals thus combined those who had no lifetime experience with any addictive substance with those who had modest to moderate exposures to legal addictive substances.

Sample 2: African American Polysubstance Abusers and Controls

African Americans who volunteered for research at the NIH IRP (NIDA) were also characterized and separated into "abusers," "controls," and samples with intermediate lifetime uses of substance use, as noted above.

Samples 1 and 2: Efficacies of Recruiting Subjects for Association Studies and Comparison with Recruitment for Linkage-Based Studies

Limited currently available data document the features of drug abuse research volunteers who might consent to participate in molecular genetic studies for linkage or association. We thus describe the details of the recruitment of subjects for these studies at the NIH (IRP) facility in Baltimore, MD, during a 30-month period (J. H., G. R. Uhl *et al.*, unpublished observations). During this period, 13,969 individuals were screened by telephone and 2633 were interviewed in person for all (genetic and nongenetic) studies at this research facility. This group included 68% African American, 29% European American, and 1% Hispanic individuals; 72% were men and their average age was 35 years.

Six hundred thirteen unrelated proband individuals from the group of interviewed subjects were offered participation in this genetic study, based on the availability of screening resources. No individual who was offered participation during this time period refused to participate. The individuals who accepted participation had an average age of 34 years and 72% were men; this group included 72% African American and 26% European American individuals as well as 2% of other ethnicities, based on self-report. Subjects accepting research participation in this study thus appear representative of this overall research volunteer population in this area. They also share some characteristics of the drug-abusing population in Baltimore, based on population trends and 1981 data from the Baltimore Epidemiological Catchment Area site, which identified 57% of the men and 41% of the women who displayed substance abuse and/or dependence as European American (J. Anthony, personal communication, 1998).

Each volunteering subject was offered three choices concerning family member contacts. One hundred twelve probands (18%) provided Type I consents, which allowed investigators to contact their family members; 73% of these individuals were men, 66% were African American, 30% were European American, and 3% self-reported other ethnicities. Three hundred twelve probands (51%) gave Type II consents, stating that they would contact their family members; 72% of these individuals were men, 74% were African American, 23% were European American, and 3% reported other ethnicities. One hundred eighty-nine (31%) refused family member contacts; 72% of these individuals were men, 74% were African American, 23% were European American, and 3% reported other ethnicities.

For 33% of the pedigrees for which the proband had provided Type I consents, at least one member could be reached by telephone or mail. At least one member of 12.5% of the pedigrees for whom the proband had provided Type II consent called an investigator and kept an appointment for study participation.

Of the 79 pedigrees from which family members made and kept appointments for study participation over this 30-month time period, 75% had African American probands and 69% had male probands. The sizes of the potentially accessible sibships in the pedigrees of these individuals, as described by the probands, were: 1 (22%), 2 (23%), 3 (13%), 4 (11%), 5 (12%), 6 (9%), 7 (7%), 8 (1%), and 11 (2%). The numbers of accessible parents from these pedigrees was 0 (10%), 1 (28%), and 2 (62%). Average families for which more than the proband were accessible thus had 1.5 accessible parents and about 3.5 accessible siblings.

Over this 30-month time period, DNA and clinical information were collected from

two-thirds of the members of the average pedigree from which any member came for an appointment. More complete sampling was obtained from smaller than from larger pedigrees. Of the 79 pedigrees for which DNA and clinical information were obtained, 54 had 2 members, 15 had 3 members, 7 had 4 members, and 3 had 5 members. By the end of this period, DNA and clinical information was successfully collected from 2.5 members of the average pedigree. Fourteen of these pedigrees contained same-gender siblings, and 15 contained opposite-gender siblings within 5 years of the age of the proband who were discordant for drug abuse phenotype. Eighteen siblings were within 5 years of age of the proband and concordant for substance abuse phenotype. Thirteen siblings differed from the probands in age by more than 5 years and were discordant for phenotype, whereas four were concordant.

Reliability of pedigree structure reports was assessed by comparing family structure information provided by the proband and another first-degree relative informant from these 79 families. Parents agreed (completely) with 100% of the proband's pedigree assignments, whereas siblings agreed (completely) with 70% of all possible pedigrees. Disagreements were largely a result of differential reporting of half versus full siblings.

Reliability of information about drug histories was assessed by comparing drug use survey (DUS) and family history/research diagnostic criteria (FH/RDC) estimates of drug use by the proband with FH/RDC estimations from first-degree relatives. Seventy-three percent of parental evaluations agreed (completely) with the proband's evaluations of all pedigree members' status with respect to all abused substances. Eighty-one percent of sibling evaluations were concordant (completely) with the proband's evaluations of his or her own drug use. Differences were most prominent in both cases as a result of family member underestimation and/or under reporting of offspring or sibling drug use.

As estimate of the validity of DUS quantity/frequency estimates was obtained by comparing DUS ratings for substances with criteria identified as strongly heritable in work by Tsuang and colleagues[102]: based on subjects' reports that the substances were "never used, used fewer than 5 times, and used more than 5 times." All individuals who denied using alcohol, cocaine, heroin, cannabis, or nicotine when questioned as part of the addiction severity index screening received "0" scores on the DUS for the appropriate substance. All but two of the individuals who scored 2+ or 3+ on the DUS reported use more than five times with a separate instrument on a separate occasion. One individual who reported use of a substance five or fewer times during screening obtained a 2+ and one obtained a 3+ score on the DUS for cannabis. Individuals who reported one to five lifetime uses during screening with the addiction severity index received 1+ DUS ratings on several occasions. The percentage of individuals who were rated as 1+ on DUS and who reported use of substances between one and five times on a different scale was 7%, 12%, 24%, 32%, and 39% for alcohol, nicotine, cocaine, cannabis, and heroin, respectively.

Although it is encouraging that none of the individuals who were offered participation in this study refused to participate themselves, only one-fifth of the probands agreed, during initial evaluations, to provide permission for research staff to directly contact other members of their families. Individuals who agreed to contact their own family members did so at rates substantially lower than the contact rates achieved when family contacts were initiated by research staff members. Such differential participation may well provide occult differences between samples collected for association versus linkage studies.

It may also be important to consider that these subjects consented to studies in which genotyping was performed using pooling approaches that provide maximal protection from research risks (see below). Samples collected in studies that propose to conduct unlimited

high-density individual genotyping in any number of different laboratories might well experience different consent rates from some subgroups of participants, providing an additional confounding factor.

Samples 1 and 2: Genotyping

The primary data reviewed here are based on assessments of allele frequencies in multiple DNA pools, each of which contained equal amounts of DNA from 20 individuals of the same racial and phenotype group. Each DNA pool was assessed on four sets of four arrays, two from 100K and two from 500K Affymetrix microarray sets. We also document some of the features of assessments using 1M SNP Affymetrix 6.0 arrays.

The 600K methods used for these samples revealed correlations ($r = 0.95$) between pooled and individual genotyping in extensive validation studies[12] (see below). Much of this variation (1%–2%) can be attributed to the variations in pipetting and DNA quantification required for pool construction (T. D. and D. W., unpublished observations, 2002–2006). The remainder of the variance is reduced for arrays used at the end of a study, in comparison to validating experiments that use data from some of the first arrays of the type that were studied within this laboratory. The correlation ($r = 0.95$) thus probably reflects an upper limit of the variance noted in actual disease versus control comparisons. Even higher correlations between individual genotyping and pooled genotyping data sets can be revealed by 1M SNP assays (Liu *et al.*, in preparation).

Data for each SNP provide a score that results in a continuous measure of the percentage of its two alleles in the DNA from the hetero- and homozygous individuals represented in each pool. For SNPs that displayed nominally significant differences between addicted and control individuals in 600K studies, the correlation between the magnitude of the differences in pooled versus individual genotyping approaches was $r \geq 0.9$. Variance from array-to-array (assessing the same DNA pool) and variance from pool-to-pool were modest, around 3%, suggesting that the validity of these pooling data is good. Results from 1M SNP assays provide even more modest pool-to-pool and array-to-array variances (Liu *et al.*, in preparation).

Sample 3: COGA European American Alcohol-Dependent Subjects and Controls[33]

Unrelated individuals sampled from pedigrees collected by the Collaborative Study on the Genetics of Alcoholism (COGA) provide an interesting sample for several reasons. Dependence on alcohol and other substances has been carefully characterized in these individuals using validated instruments. Unrelated control individuals free from substance abuse or dependence diagnoses, largely individuals who marry into these pedigrees, are available. We thus identified 120 unrelated alcohol-dependent individuals and 160 unrelated unaffected controls with self-reported European American ethnicities.

Sample 3: Genotyping

Allele frequencies were assessed in 14 DNA pools, each containing equal amounts of DNA from 20 individuals of the same phenotype group, using four sets of four arrays, two from 100K and two from 500K Affymetrix SNP arrays, using approaches that were extensively validated as noted above.

Sample 4: Taiwanese Methamphetamine-Dependent Subjects and Controls

Unrelated subjects recruited in Taipei included 140 methamphetamine-dependent individuals independently diagnosed by each of two psychiatrists using *DSM–IV* criteria[103–108] and 240 matched Han Chinese controls who

denied any history of use of illegal drugs and denied any history of psychotic symptoms. Thirty percent of the subjects were women and their average age was 32.5 ± 10 years. Dependent individuals reported methamphetamine use more than 20 times per year or described well-documented methamphetamine psychosis with lower levels of regular use. They denied histories of psychosis either prior to methamphetamine use or in relation to other psychedelic drugs. Most reported use of at least one other addictive substance. Controls denied illegal drug use or psychotic symptoms and were matched for gender and age.

Sample 5: JGIDA Japanese Methamphetamine-Dependent Subjects and Controls

Twenty-one percent of the Japanese subjects were women and their average age was 40 years. One hundred methamphetamine-dependent subjects were inpatients or outpatients of psychiatric hospitals in the regions that participate in the Japanese Genetics Initiative for Drug Abuse (JGIDA)[12,104–116] and met ICD-10-DCR criteria F15.2 and F15.5 for methamphetamine dependence in independent diagnoses made by each of two trained psychiatrists based on interviews and review of records. Ninety-one percent revealed histories of methamphetamine psychosis, 89% used methamphetamine intravenously, 62% also abused organic solvents, and most abused at least one other substance. Subjects who displayed clinical diagnoses of schizophrenia, other psychotic disorders, or organic mental syndromes were excluded. Control subjects included 100 age-, gender-, and geographically matched staff recruited at the same institutions who denied use of any illegal substance, abuse of or dependence on any legal substance, any psychotic psychiatric illness, or any family history of substance dependence or psychotic psychiatric illness during interviews with trained psychiatrists.

Samples 4 and 5: Genotyping

We assessed allele frequencies for methamphetamine-dependent and control subjects in DNA pools, each containing equal amounts of DNA from 20 individuals of the same phenotype group, on four sets of arrays, two arrays from 100K and two from 500K Affymetrix sets (Sample 5) or two arrays from 500K Affymetrix sets (Sample 4).

Sample 6: Australian and U.S. Dependent versus Nondependent Smokers of European Ancestry[3]

Dependent smokers of European ancestry were diagnosed using the Fagerström Test for Nicotine Dependence criteria (FTND score ≥ 4) and were compared with smokers who did not display dependence (FTND scores $= 0$).[3] About one-quarter of the individuals who displayed dependence by FTND standards did not display *DSM–IV* nicotine dependence. About one-quarter of the individuals used in this study as controls did display DSM nicotine dependence.

Sample 6: Genotyping

Allele frequency data were assessed in 16 DNA pools, each containing equal amounts of DNA from 60 individuals using a single set of 49 arrays that assessed 2,427,354 SNPs. Methods used for these samples revealed overall correlations ($r = 0.85$) between pooled and individual genotyping for all SNPs. However, much more modest correlations ($r = 0.58$) were found between dependent versus nondependent data derived from pooled versus individual genotyping for the SNPs that displayed nominally significant differences. Individual genotyping followed up nominally positive results for 39,213 SNPs in 1050 dependent and 879 nondependent smokers. The convergence analyses presented here use data from the subset of these 39,213 SNPs that lie within genes.

Sample 7: WTCCC Subjects with Bipolar Disorder and Controls

GWA for bipolar disorder compared controls to 1868 U.K. individuals of European descent with bipolar disorders from the Wellcome Trust Case Control Consortium (WTCCC).[1] Bipolar mood disorders were diagnosed using Research Diagnostic Criteria. Uncharacterized control samples included (a) 1480 individuals from a 1958 birth cohort sample, (b) 1458 individuals from a UK Blood Service sample of consenting blood donors, and (c) individuals with disease phenotypes whose genetics were deemed unlikely to overlap with the genetics of bipolar disorder.[1]

Sample 7: Genotyping

Genotyping for the 436,604 autosomal SNPs analyzed was performed using Affymetrix 500K arrays with allele calls made by a CHIAMO algorithm with an *a posteriori* probability threshold of at least 0.9.[1] Of the 469,557 SNPs assessed, 436,604 could be assigned confident chromosomal localizations. A *P* value for each SNP was determined based on its χ^2-test for significance of allele frequency differences in bipolar versus control (i.e., control plus other disease) subjects.[1]

Sample 8: NIMH Subjects with Bipolar Disorder and Controls

GWA was assessed in controls compared with 461 unrelated bipolar I probands of self-reported European American ancestry who were selected from families that included at least one affected sibling pair who participated in the National Institute of Mental Health (NIMH) Genetics Initiative (http://nimhgenetics.org).[2] Probands were assigned a "confident" diagnosis of *DSM–IV* bipolar I disorder by each of two trained clinicians. Five hundred sixty-three unrelated control individuals of European American ancestry who failed to display evidence for *DSM–IV* criteria for major depression, any history of bipolar disorder, or any history of psychosis were recruited by a marketing firm.

Sample 9: German Subjects with Bipolar Disorder and Controls

GWA for 536,288 autosomal SNPs was performed in controls compared with 772 bipolar I patients diagnosed using *DSM–IV* criteria who were recruited from consecutive hospital admissions.[2] Eight hundred seventy-six population-based controls were randomly recruited; individuals with histories of affective disorder or schizophrenia were excluded.

Samples 8 and 9: Genotyping

For genotyping, NIMH samples (Sample 8) were divided into seven bipolar and nine control pools of 50–80 subjects per pool. German samples (Sample 9) were divided into 13 bipolar and 10 control pools of 42–60 subjects per pool. SNP allelic distributions were assessed using duplicate Illumina HumanHap550 assays (Illumina Inc., La Jolla, CA, USA).[2] Normalized allele frequencies were calculated from raw intensity data averaged across duplicate pools to obtain a relative allele frequency estimate for each SNP in each pool. SNPs with allele frequencies that displayed greater than 2% variance between replicate pools were excluded. Pool-to-pool variation within phenotypes was compared with phenotype-to-phenotype differences using *t*-tests.

Sample 10: Unrelated Members of NHLBI Twin Pairs for Assessment of Frontal Brain Volume

Two hundred forty-two unrelated individuals were selected randomly from members of twin pairs from a population-based registry of European American male World War II veteran twin pairs who received volumetric MRI studies as part of the National Heart, Lung, and Blood Institute (NHLBI) Twin Study. When studies were performed, subject age averaged

72.6 years, and they reported an average of 13.6 years of education. Frontal lobar volumes, corrected for intracranial volumes, were obtained[117] in ways that produced interrater reliabilities greater than 0.90.

Sample 10: Genotyping

For genotyping, DNA samples were carefully quantitated and combined into 12 pools that each represented about 20 subjects based on estimates for frontal brain volume corrected for total cranial volume. We thus constructed four DNA pools from individuals with the lowest estimates of total frontal brain volumes, four pools from individuals with intermediate volumes, and four pools from individuals with the highest estimates of total frontal brain volumes. We subjected these DNA pools to Affymetrix 500K genotyping as noted above and used *t*-tests to compare differences between the highest and lowest brain volume groups. SNPs that displayed nominally significant *t* values and that also displayed rank order of allelic frequencies with either highest tercile > intermediate tercile > lowest tercile or the converse are included in these analyses.

Sample 11: Framingham Study Participants for Assessment of Frontal Brain Volume

Subjects were 705 stroke- and dementia-free participants in the Framingham study. The average age of subjects was 62 $\pm$ 9 years, and 50% were men. Subjects received volumetric brain MRI studies that were analyzed as noted above for the NHLBI subjects.

Sample 11: Genotyping

Genotyping provided data from 70,987 autosomal SNPs using Affymetrix 100K arrays. Allele frequencies for SNPs that displayed (a) minor allele frequencies ≥ 0.10, (b) genotype success ≥ 0.80, and (c) Hardy–Weinberg equilibrium $P \geq 0.001$ were used. A generalized estimating equation provided corrections for familial relatedness and other covariates.[8]

Sample 12: European American Smokers Who Successfully versus Unsuccessfully Quit Smoking in Trials in Philadelphia, PA, Washington, DC, and Buffalo, NY

European American smokers who successfully versus unsuccessfully quit smoking in trials in Philadelphia, Pennsylvania, Washington D.C., and Buffalo, New York responded to advertising and physician referrals for help in smoking cessation.[118,119] Subjects aged 18–65 enrolled in randomized clinical trials for smoking cessation accompanied by standardized behavioral counseling that used a blinded, placebo-controlled trial of bupropion (300 mg/day) or matching placebo for 10 weeks or an open-label trial of nicotine nasal spray versus nicotine patch for 8 weeks.[118] One hundred twenty-six individuals with biochemically confirmed abstinence for at least the 7 days prior to assessments at both 8 weeks and 24 weeks were contrasted with 140 unsuccessful quitters who were not abstinent at either time point.

Sample 13: European American Smokers Who Successfully versus Unsuccessfully Quit Smoking in Trials in North Carolina

Participants received either active nicotine (21 mg/day) or placebo skin patches for two weeks before the targeted quit date as well as mecamylamine (10 mg/day p.o.) prior to the target quit-smoking date.[120] After the quit date, participants were randomly assigned to mecamylamine (10 mg/day) versus matching placebo, and to 21 mg/24 h versus 42 mg/24 h nicotine skin patch doses. Fifty-five individuals reported continuous abstinence from smoking when assessed 6 weeks after the quit date with biochemical confirmation; 79 were not abstinent.

Sample 14: European American Smokers Who Successfully versus Unsuccessfully Quit Smoking in Trials in Rhode Island

Participants engaged in a 10-week, double-blind, placebo-controlled trial of placebo or bupropion (150 mg/day for the first 3 days, then 300 mg/day) with a target quit date 1 week following initiation of drug or placebo.[119] Sixty individuals with biochemically confirmed abstinence for at least the 7 days prior to the end of treatment and at a 24-week assessment were contrasted with 90 unsuccessful quitters who were not abstinent at either time point.

Samples 12–14: Genotyping

Genotyping for samples 12–14 used Affymetrix 500K arrays and multiple pools of DNA samples ($n = 16$ to 20), as noted above.

Sample 15: African American Individuals with Different Levels of General Cognitive Ability

Research volunteers with a variety of different levels of lifetime use of addictive substances volunteered for research protocols at the NIH (NIDA) facility in Baltimore, Maryland, as described above. Eighteen pools were constructed with DNA from 20 individuals each; 33% of the subjects were women, and the average age was 32.1 years (range: 18–65 years). Mean cognitive function scores estimated from the Shipley Institute of Living scales ranged from "IQ" equivalents of 75.9 to 109.2 for the individuals in these DNA pools.

Sample 16: African American Individuals with Different Levels of General Cognitive Ability

Eleven pools were constructed with DNA from 16 individuals each; 34% of subjects were women, and the average age was 31.8 years (range: 18–65 years). Mean estimated "IQ" scores from the Shipley Institute of Living scales ranged from 79.1 to 112.3 for the individuals represented in these DNA pools.

Samples 15 and 16: Genotyping

Genotyping for samples 15–16 used Affymetrix 500K arrays and pools of DNA samples ($n = 16$ to 20), as noted above. The nominal significance of the correlations between pool-to-pool differences in assessments of allele frequency and pool-to-pool differences in Shipley scores was assessed for each SNP.

Sample 17: Subjects with Alzheimer's Disease versus Controls: Brain Donors

Subjects were 1086 brain donors who were at least 65 years old at death (with a mean age at death of 82 years); 43% of subjects were men. Brains and clinical data met pathological criteria for Alzheimer's disease (AD) or control status. DNA samples were subjected to Affymetrix 500K genotyping. From the files TGEN_WGA_DATA_ recode_ped.txt,[4] genotype calls for 552 control individuals and 859 individuals with AD allowed us to (a) calculate χ^2 values for the AD versus control differences for each SNP, (b) select SNPs that displayed χ^2 values with $P < 0.05$ as nominally positive, and (c) assess which of these nominally positive SNPs fell into chromosomal clusters such that at least three nominally positive SNPs representing both array types lie no more than 25 kb from each other.

Sample 18: Subjects with Alzheimer's Disease versus Controls: Memory Clinic Participants

Seven hundred fifty-three individuals with AD and 736 control subjects with European ancestry were recruited in Canadian memory clinics. Probable AD was diagnosed by clinical criteria, and controls were selected who displayed no histories of memory impairment or any impairment on neuropsychological tests.[5] DNA samples were subjected to Affymetrix

500K genotyping. From files available through the GlaxoSmithKline Clinical Trial Register (available at http://ctr.gsk.co.uk/Summary/genetic_observational/studylist.asp), P values for each SNP, derived from Fisher's exact tests, were extracted and data were analyzed as described above.

Sample 19: Individuals with Scores on Tests of Neuroticism

One thousand thirty-eight individuals from southwestern England sites with European backgrounds and with high neuroticism (N) scores on the revised Eysenck Personality Questionnaire and 1016 individuals with low N scores were studied; 63% of these subjects were women. A replication sample (61% female) included 831 high versus 702 low N individuals.[121] Genotyping of eight pools of DNA from mouth swabs compared (a) men with high N scores ($n = 112$), (b) men with low N scores ($n = 158$), (c) men with very high N scores ($n = 245$), (d) men with very low N scores ($n = 238$), (e) women with high N scores ($n = 320$), (f) women with low N scores ($n = 205$), (g) women with very high N scores ($n = 340$), and (h) women with very low N scores ($n = 436$). Very high or low N scores were defined as more than 1.5 SD from the mean score adjusted to age and sex (on average, 2 SD), whereas high and low N scores were between 1 and 1.5 SD from the mean score (on average, 1.3 SD). Data for relative allele score from the 452,574 SNPs from 100 and 500K Affymetrix arrays with minor allele frequencies above 5% were obtained from five replicate arrays used to assess each sample.

Selected Methodological Issues

Genotyping Using Pooled DNA

Genotyping for GWA studies can be performed in either individual samples or in pools of DNA from individuals with the same racial or ethnic background and the same phenotype. Pooling strategies have several advantages. Pooling fits well with association genetics, can allow for efficient allele typing, can preserve confidentiality, and can reduce costs.[10,122–124]

Not all pooling strategies are alike. *Single-pool* strategies seek differences in allele frequencies by comparing data between a single pool of DNA from diseased individuals and a single pool of DNA from control individuals. Such results generate hypotheses. However, such designs, and related designs with very few pools, provide little ability to differentiate between (a) the variability between disease and control samples and (b) the variability within disease samples or within control samples.

We focus here instead on *multiple-pool* strategies. With careful attention to a large number of small details, these approaches can provide accurate allele typing. Multiple-pool approaches provide estimates of (a) differences between disease and control samples, (b) variability within disease samples, and (c) variability within control samples. Assessment of the differences between disease and control in the context of assessments of the variability within disease samples and within control samples allows us to use standard statistical approaches to assess the significance of the results.

Here, we provide assessment of several of the steps necessary to validate and characterize features of the power and sensitivity of multiple-pool genome-scanning strategies. Used with high densities of genomic markers, carefully performed multiple-pool studies can provide increased study feasibility and preserve virtually absolute genetic confidentiality with only modest effects on the sensitivity and specificity of GWA.

(a) ***DNA quality, quantity, and contamination.*** Care in assessing and maintaining the quantity and quality of DNA in every sample is crucial for pooling studies. Rough DNA quantitation procedures that are routinely used in most genotyping laboratories are likely to introduce such substantial errors that many of the apparent

disease versus control differences will actually arise from occult over- or underrepresentation of genotypes of individuals with misquantitated DNA. Uneven DNA quality can provide the same selective over- and underrepresentation of genotypes of selected individuals in each pool, leading to more false positives and less sensitivity for detection of true positive results. Contamination of pooled DNA samples with even a modest amount of DNA from laboratory personnel or other sources can also provide difficulties for pooling procedures.

(b) ***Numbers of individuals per pool and numbers of pools.*** To obtain maximal benefits from multiple-pool GWA, one must meet the following requirements: (a) The numbers of individuals in each pool should be sufficient that even sophisticated analyses of pooled data cannot reconstruct individual identities or genotypes. Treatments of this subject suggest that pools need to contain more than four or five individuals for maximal confidentiality protection.[125] (b) The numbers of individuals in each pool should allow significant cost and time savings compared with individual genotyping. (c) The numbers of pools should be sufficient to provide good estimates of pool-to-pool variability, which can then be used to compare to the differences between disease and control individuals using standard statistical tests.

Multiple genotyping assessments of each DNA pool can aid the precision of estimates of pool-to-pool variability as well. We have used three to four microarrays to assess DNA samples from each pool. These numbers are based on preliminary studies that seek to optimize estimates of "true" relative allele frequencies at acceptable cost. We construct each pool using DNA from 20 individuals of the same self-reported ethnicity and the same disease or "control" phenotype. Thus, we obtain results in quadruplicate at 1/5–1/7 the reagent costs of individual genotyping using a single array set per person.

(c) ***Number of different replicate samples.*** Power calculations assess the likelihood that an experiment can detect a difference of a certain magnitude in a specific SNP. Experiments require reasonable levels of protection against false positive results, α. They also require reasonable power, β. GWA requires many repeated measures; considerations of α and β thus need to be applied to hundreds of thousands or millions of SNPs.

One approach to the dilemma raised by these large numbers of multiple comparisons has been to propose single studies with increasingly large sample sizes. However, accretion of very large samples is expensive. Attempts to assemble large samples from smaller subsamples collected at various sites also run greater and greater risks of incorporating increasing numbers of occult heterogeneities that could provide confounding influences on the results obtained.[126]

The approaches that we outline here rely on initial use of achievable sample sizes that may be more likely to be more homogeneous. Initial samples can nominate sets of SNP markers, genomic regions, and genes that can be studied in additional independent, replicate samples. A requirement that genes display SNPs whose allelic frequencies distinguish disease from control individuals in multiple samples is one of the few assurances against false positive results that is likely to pass ultimate statistical muster and also to yield feasible experimental designs. There is a downside to this approach: rates of false negative results are also unavoidably elevated by requirements for replication (see below).

(d) ***Validation studies.*** Validation studies assess the fits between individual and pooled genotyping in a number of different ways. These include assessments of the concordance between results from sense and

antisense probes for the same SNP and concordance between results for the same SNPs obtained using arrays of different types. The same DNA samples can be pooled multiple times, and the same pool can be analyzed multiple times, to further assess concordance.

We focus here on a core validating test for pooling. This core test comes from analyses of the relationships between (a) observed allele ratios, background-subtracted, normalized hybridization intensity ratio values obtained from different pools of DNA samples; and (b) expected allele ratios, the fraction of, for example, "A" and "B" alleles obtained from individual genotypes. We and others have compared allelic determinations from individual DNA samples versus results from pools with equal or differing amounts of DNA from small numbers or larger numbers of control individuals using HuSNP; Affymetrix 10K, 100K, 500K, and 1 million SNP products; Perlegen arrays; and Illumina 300 and 500K arrays.[11,33,39,41,42,122,123,127]

Using 500K Affymetrix arrays, we have evaluated pooling using equal and varying amounts of DNA from CEPH individuals. Overall data for 150,000 SNPs from these comparisons produces correlations between pooled and individually determined genotypes of $r^2 = 0.95$ (Fig. 3). These overall results derive from studies of 5475 and 6230 informative *Nsp* I and *Sty* I SNPs in experiments in which equal amounts of DNA from homozygotes were mixed (correlation ca 0.9); 10,032 informative *Nsp* I and 10,249 informative *Sty* I SNPs in experiments in which these same DNA samples were mixed in 1:1, 1:5, and 1:15 ratios (correlations of 0.96 and 0.98, respectively) and 31,201 informative *Nsp* I and 39,827 *Sty* I SNPs for studies of one homozygote and one heterozygote (correlations of 0.89 and 0.92, respectively). When we compare these results with those reported using long-range PCR products, using the reported procedure of eliminating the 9% of SNPs that yielded the more problematic correlations,[127] the overall correlations for the remaining SNPs is 0.98. Correlations using 1M Affymetrix SNP arrays are at least as strong, with $r^2 > 0.98$.

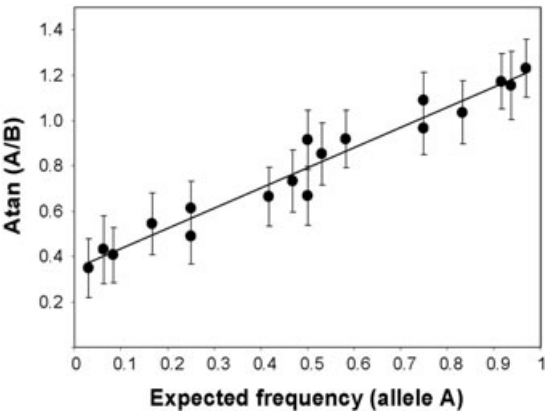

Figure 3. Validation of SNP genotyping in DNA pools (From 12). The relationship ($r = 0.95$) between individual and pooled genotyping using 500K SNP Affymetrix arrays provides an opportunity to assess the sensitivity of pooled genotyping. Because these validation experiments were the first ones performed with new array sets, these data provide a lower limit. Current results from 1M SNP arrays (6.0) provide relationships of around R = 0.98 (Drgon *et al.*, in preparation).

Power Assessments

Approaches to assessing power of GWA have used a variety of assumptions about the frequencies of disease-causing alleles, the heterogeneity and penetrance of disease-causing alleles, marker frequencies, and the nature and distribution of linkage disequilibrium across the genomic intervals surveyed.[128,129] Many approaches to this problem use linkage disequilibrium distributions identified in HapMap samples, even though these HapMap individuals represent only very small subsets of several current human populations.

We have been impressed by the variability in the detailed distribution of linkage disequilibrium across different genomic loci in different

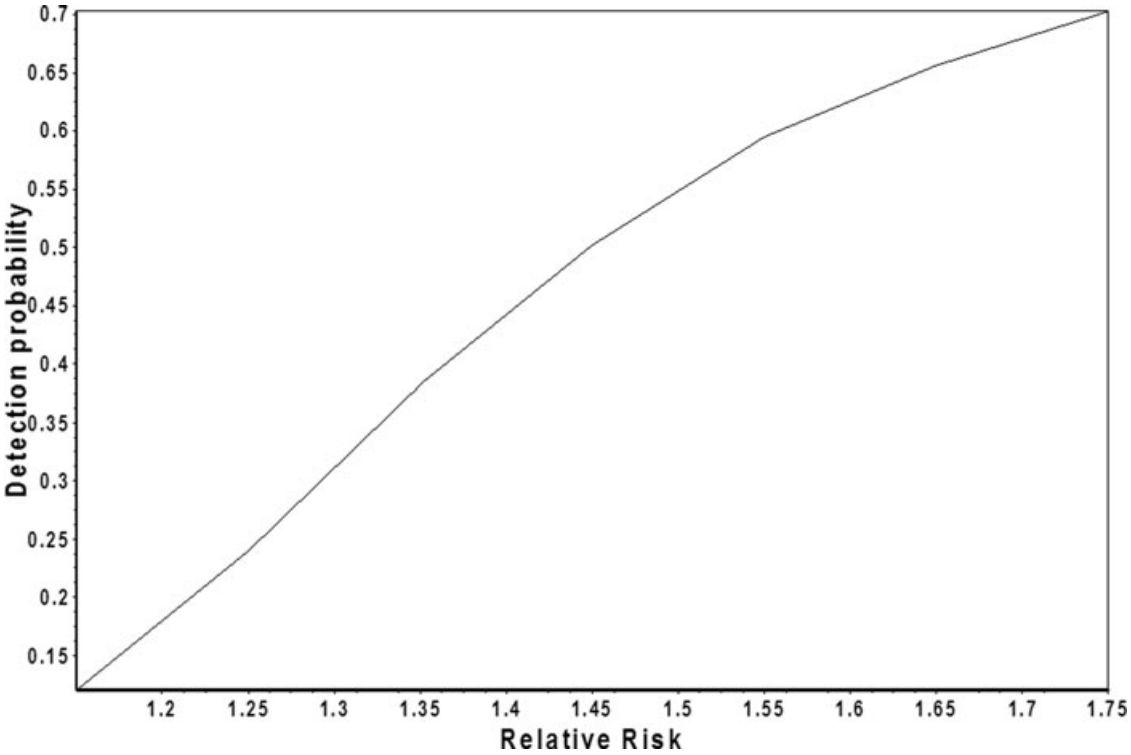

Figure 4. The power of GWA as assessed using Gene Detective. Power is simulated here with 620,000 diallelic markers for samples with $n = 400$ cases and $n = 400$ controls with nominal 0.05 α levels. Note the striking relationship between power and effect size. The power to detect effects that would produce odds ratios of less than 1.2-fold is modest, whereas the power to detect effects as high as 1.7-fold is relatively good.

samples.[130,131] We have also been impressed by the potential to model approximations of this variability, on average, using simple functions. Under these circumstances, estimates of effects of sample size, locus-specific effect sizes for the underlying functional alleles, genetic heterogeneity, penetrance, and marker density can produce reasonable models that can allow assessments of the effects of variation in these parameters on power.

We have focused on diallelic markers and disease/no disease phenotypes. We have developed a model to simulate the effects of varying these parameters that has resulted in the program Gene Detective. We can use this approach (see Supplement for details) to simulate the approximately 620,000 diallelic markers reported to date for samples with $n = 400$ cases and $n = 400$ controls with nominal 0.05 α levels (Fig. 4). We can observe effects of sample size, heterogeneity/penetrance ratios, marker minor allele frequencies, and disease frequencies. Such effects are each relatively modest over a reasonable range of values for genome-wide distributions of linkage disequilibrium. However, the relationship between power and effect size is striking. The power to detect effects that would produce odds ratios of less than 1.2-fold is modest, whereas the power to detect effects as high as 1.7-fold is relatively good.

Increments in marker density from 630,000 to 1,000,000 and increases in sample sizes from $n = 400$ to $n = 2000$ samples in case and control improve power (Fig. 5). However, the steep relationships between power and effect sizes are also found in these simulations. Such

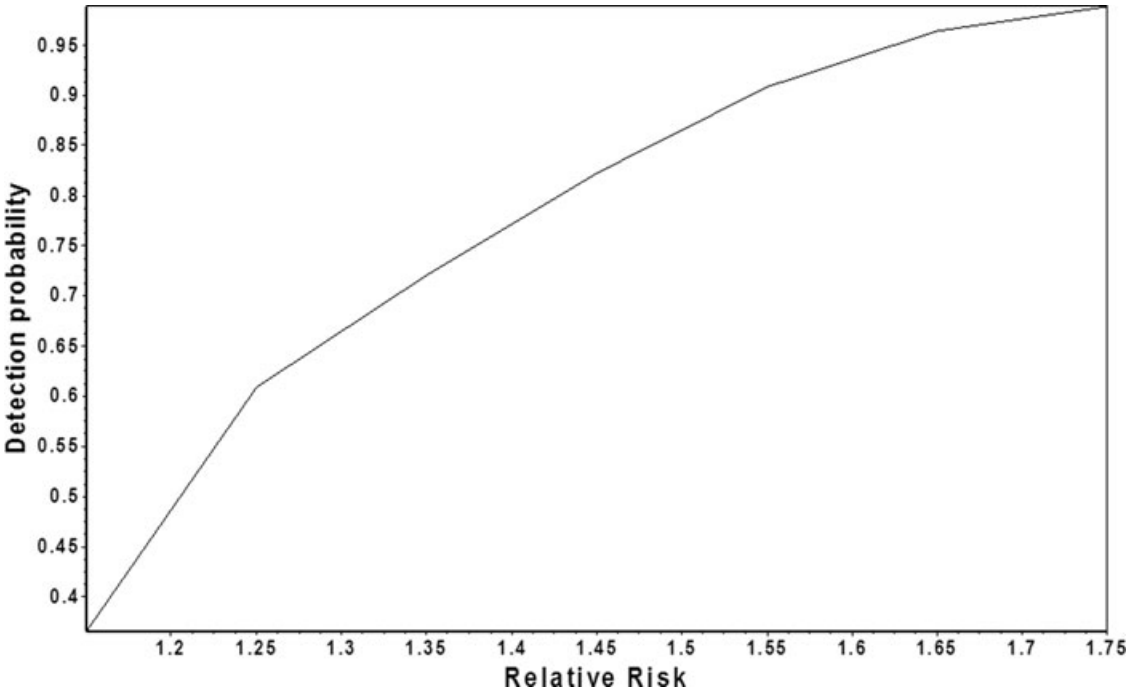

Figure 5. The power of GWA as assessed using Gene Detective. Power is simulated here with 1,000,000 diallelic markers for samples with $n = 2000$ cases and $n = 2000$ controls with nominal 0.05 α levels. Note that the striking relationship between power and effect size is retained.

results underscore the distinctions that we have made above concerning analytic approaches to oligogenic disorders, in which variants at individual gene loci produce relatively large differences in risk, versus polygenic disorders, in which the effects at each locus are likely to be modest.

These power calculations apply only to the initial genome scans. As we note above, replicate studies aid in distinguishing false positive from true positive associations but also increase the cumulative number of false negative results.

Limits on the precision of the power calculations that derive from the approach outlined here include limits on the precision of estimation of the parameters whose estimation is required. Several of these parameters can be estimated based on substantial empirical data. These include the size of the genome or genomic segment under consideration, the sizes of the samples of disease and nondisease control subjects studied, and the value for α desired. The frequency of the disease, P [D], in the population under study is available from epidemiological studies. It is thus important that the samples for association genome scanning display characteristics similar to those of the populations in which disease probabilities have been determined, so that estimates of P [D] are as accurate as possible.

To assess the statistical power of analysis, we have also used more standard power calculations. The program PS v2.1.31[132] (a) $\alpha = 0.05$, (b) sample sizes equal to the numbers of pools from the current data set, (c) mean abuser/control differences of 0.05 and 0.1, and (d) standard deviations from the SNPs that provided the largest differences between control and abuser populations are used in several of the discussions below. We have also used data from the Genetic Power Calculator (http://pngu.mgh.harvard.edu/~purcell/gpc/) for some analyses.

Analyses

Achieving Significant Genome-Wide Association in Single Samples versus Seeking Replication and Generalization in Multiple Samples

(a) ***Single-sample approaches.*** As noted above, GWA gains power to detect variants in more and more of the genome as more and more genetic markers, generally SNPs and/or copy number variants, are assayed. Because many hundreds of thousands of SNPs and/or copy number variants are assayed in current data sets, stringent approaches to correct for the large number of multiple comparisons are needed.

No clear-cut consensus has been reached regarding the ability of any single method to produce only true results from any single sample. One approach to concerns about the large numbers of comparisons that are key components of GWA focuses on achieving *genome-wide significance in single samples*. Single samples that demonstrate genome-wide significance in this way must contain single SNPs whose association displays a striking nominal *P* value, often in the neighborhood of about 10^{-8}.[133,134] Such results may be the most likely to be published in prominent journals. However, in most studies with findings of this statistical magnitude, effects of variants at a single locus are sufficiently large that linkage studies also provided significant evidence at the same chromosomal locus.[135,136] For oligogenic contributions to common, complex disorders, seeking association with genome-wide significance in single samples thus provides a reasonable approach. When a single gene has a large effect, a number of corrections for multiple comparisons can be applied without creating many false negatives. Bonferroni corrections for multiple comparisons are advocated by some investigators, although they are generally considered to provide a conservative correction.[137,138] False discovery rate corrections can also be applied.[139–141] Permutation and Monte Carlo tests provide additional approaches.[6,142,143]

As the expected effect of each locus falls from the large effects characteristic of oligogenic influences to the small effects that characterize polygenic influences, however, the sample sizes needed to generate *P* values in these ranges provide a daunting problem. Costs of individually genotyping such large samples become limiting in all but the best-supported enterprises.[144] The risks of introducing occult heterogeneities increase when subsamples are collected at a variety of distinct sites.[126] As more occult heterogeneities are included—because disease and control samples need to be assembled from more and more diverse sources to achieve a sufficient sample size—more and more of the results obtained may well represent false positives based on such occult sample heterogeneity for genetic background or for heritable traits that are not (nominally) being studied.

(b) ***Replicate sample approaches.*** Here we use an alternative analytic approach that focuses on stepwise assessments. These stepwise analyses address the problem of multiple testing by seeking nominally significant results that can be replicated in several independent samples. We can assess the significance of these replicated, nominally significant results through the use of Monte Carlo methods that correct for multiple comparisons.

It is important to emphasize the ways in which the stepwise analyses presented here identify, first, evidence for genes that contain haplotypes found at different frequencies in single disease versus control sample comparisons and, second, evidence for genes that display haplotypes with such different frequencies in multiple samples that results are unlikely to be due to chance.

We thus (a) first identify nominally significant SNPs in each sample, (b) identify the clustering of such SNPs (within small

chromosomal regions) in each sample, (c) seek replication, identifying small genomic areas in which clusters from multiple replicate samples from the same phenotype also identify clustered nominally significant SNPs, and (d) seek generalization, identifying genes that contain clusters of nominally positive SNPs from studies of related, genetically determined phenotypes.

The criterion used here identifies clustering based on chromosomal position. This approach allows direct comparison between data sets that assess different sets of SNPs in samples that may well differ in the details of their patterns of linkage disequilibrium. The Monte Carlo simulation methods used here do not make assumptions about the underlying distribution of the data assessed. Monte Carlo methods provide empirical *P* values based on repeated random samples from the actual data sets analyzed. Such approaches are especially useful when we seek to assess the significance of apparently reproducible results from convergent data from multiple independent data sets that differ from each other in sample size, number and types of genomic markers, racial or ethnic background of the subjects, and other key features. No alternative method of which we are aware provides as tractable a method for assessing the significance of results obtained in multiple samples without assumptions about underlying distributions of the data as do Monte Carlo approaches. We use 10,000 Monte Carlo trials in circumstances in which moderately high significance is anticipated, and 100,000 trials in circumstances in which extremely high significance is anticipated.

This approach seeks to identify genes with variants that are likely to play roles in addiction and in related phenotypes. This approach allows for locus heterogeneity and thus does not use the more stringent criterion that the same SNP is required to display nominal significance in each of the samples in which association data are said to support association at a specific gene locus. This approach allows for differences in the phase of association and thus does not use the more stringent criterion that the same allele of the SNP (or haplotype) must be associated with nominal significance in each of the samples in which association data are said to support association at a specific gene locus. The approach allows for different details of the patterns of linkage disequilibrium between marker and functional haplotype from sample to sample. The approach allows us to combine data sets in which different marker sets are used. With each of these limitations, it is clear that subsequent follow-up analyses are required. Analyses in the same and in additional independent samples are required to untangle any locus heterogeneity, to unequivocally identify which individual SNPs are associated, and to identify pathological haplotypes and the phases with which they are associated with phenotypes in samples from different racial and ethnic backgrounds. Although we describe examples of such follow-up studies for the NrCAM and neurexin 3 (NRXN3) genes below, it is important to note the limited numbers of genes for which such confirmatory follow-up data are available. In many circumstances, we believe that this sort of follow-up requires molecular biologic, behavioral, and other evidence to buttress the data that come from association genetics alone.

Stepwise Approaches to Analyses

(a) ***Determination of nominally significant markers.*** Nominal *P* values that come from *t*-tests (for pooled data), χ^2 (for individual genotype frequencies), or ρ (for correlational approaches) statistics delineate the nominal significance of the differences between disease and control groups for each SNP. For pooled assessments, proper definition of the pool-to-pool variability is crucial for proper assignment of the appropriate nominal *t* value.

However, the continuous results that come from pooled data sets do provide the additional statistical power characteristic of statistics based on continuous measures.

(b) ***Identifying chromosomal clusters of nominally significant markers in single samples.*** We focus on the SNPs whose chromosomal positions can be accurately determined. Because gender ratios differ substantially in many of these data sets, we omit data from sex chromosomes for most of these samples.

We focus on clusters of nominally positive autosomal SNPs that lie within 100 or 25 kb of each other, depending on the density of markers available; the latter figure is closer to the average haplotype block length in the samples studied here.[11,12,33,34,36,37] To use a valuable technical control that is possible with Affymetrix 500K reagents, we require that SNPs in each cluster come from both *Sty* I and *Nsp* I array types where possible.[12] In assessment of the data from each sample set, these criteria thus provide some assurance that haplotypes do occur at differing frequencies in disease versus controls. These criteria provide significant technical controls, based on requirements that multiple nearby SNPs must display positive results and that these positive results must come from two array types.

It is important to note that, if stochastic events produce a nominally significant association at a given SNP in a single sample, linkage disequilibrium with nearby SNPs might provide a cluster of several SNPs with nominal significance in this single sample on stochastic grounds alone. Control for the possibility that these differences in haplotype frequencies result from stochastic differences between samples thus awaits the next analytic step (c, below).

We test the nonrandomness of clustering of nominally significant SNPs using Monte Carlo simulations. We can also use these approaches to identify the nonrandomness of clustering within genes. For each simulation trial, a random set of SNPs from the database that contains the results from these studies is subjected to the same analytic procedures that had been used for the actual data analysis. The number of trials for which the results from the randomly selected set of SNPs match or exceeded the results actually observed from the SNPs identified in the current study is tabulated. Empirical *P* values are calculated by dividing the number of trials for which the observed results are matched or exceeded by the total number of Monte Carlo simulation trials performed. This method examines the properties of the actual SNPs contained in each data set. It is therefore relatively robust despite the uneven distributions of SNP markers across the genome, differences in linkage disequilibrium across the genome in different samples, and the different SNPs genotyped using different assays.

(c) ***Identifying the clustered, nominally positive SNPs with the strongest positive support from several* replication *data sets.*** We next seek convergence between data from several replicate samples. We focus on samples that test the same underlying hypothesis (i.e., that common allelic variants contribute to genetic components of vulnerability to develop substance dependence). Some of these samples and their matched controls differ from each other on other bases (e.g., racial or ethnic background or primary substances of abuse). We thus use *replication* here in a restricted sense that allows us to reserve use of the term *generalization* to denote identification of genes whose pleiotropic influences are evident in studies of other heritable phenotypes that often co-occur with addictions (see below). Obviously, an aspect of generalization also applies when comparing data from (a) polysubstance-dependent versus control samples collected from individuals of two racial or ethnic backgrounds with (b) methamphetamine-dependent versus control samples collected from

individuals with a third racial or ethnic background (see below).

Analyses focus on genes identified by clustered positive results from several samples. This approach, rather than a focus on individual SNPs whose informativeness might differ across samples, allows for some degree of genetic heterogeneity and for some sample-to-sample differences in the detailed patterns of linkage disequilibrium.

Clustering of positive results in the same gene in each of several independent samples is much less likely to represent purely stochastic effects than observations made in any single sample. Such clustering in multiple samples is more likely to reflect true differences related to the phenotype of interest, such as differences in terms of a dependence on addictive substances. However, it is important to emphasize again that these criteria are aimed at the identification of genes, rather than a precise definition of exact disease-associated haplotypes. We thus allow the phase of association to differ between samples at this level of analysis. Detailed studies of the phase of association can provide a very valuable fine mapping tool to allow for the identification of the exact pathogenic haplotype.[131]

(d) ***Identifying the clustered, nominally positive SNPs with the strongest positive support from several* generalization *data sets.*** To seek possible generalization of results, we have sought chromosomal locations where the clustered positive data from several substance-dependence GWA samples lie near clustered, nominally positive (and reproducibly positive) results from studies of other related, heritable phenotypes.

Bayesian approaches to these analyses suggest that the stronger the evidence for co-heritabilities of substance dependence and these related phenotypes, the higher the likelihood that molecular genetic studies will demonstrate true overlaps.[145] We focus on phenotypes that display good evidence for heritability from classical genetic studies, including evidence that complex genetics plays substantial etiologic roles. We focus first on heritable phenotypes that co-occur with addictions at frequencies much greater than those that we would expect if they were independent of each other. For example, even though substance dependence and bipolar disorder are both common, the product of their population frequencies does not nearly explain the approximately two-thirds of bipolar individuals who report abuse of or dependence on an addictive substance.[146,147]

Twin data that compare co-occurrence frequencies in monozygotic versus dizygotic twin pairs provide evidence for shared heritability of some of these phenotypes. For other phenotypes, the magnitude of genetic influences and the frequency of co-occurrence with substance dependence indicate the likelihood of pleiotropic influences of some of the same allelic variants on both phenotypes. Finally, we also present here the idea that *transitive* genetic approaches may also identify evidence for generalization of effects of some pleiotropic influences. If substance dependence and intermediate heritable phenotypes share genetic overlap and co-occur, then a third heritable phenotype that is documented to co-occur with the intermediate phenotype might also share substantial heritability with addiction vulnerability.

Examples of heritable phenotypes for which twin data document shared genetic determinants include frontal lobe brain volume and cognitive abilities.[15,117] Examples of heritable phenotypes for which co-occurrence makes shared genetics highly likely, *a priori*, include substance dependence and bipolar disorder.[148,149] A transitive genetic approach could be applied to the shared genetics of substance dependence and cognitive abilities or brain volume on one hand, and the likely shared genetics of cognitive ability or brain volume and

vulnerability to AD, on the other hand. Data from cognitive function and frontal brain volume genetics thus provide potential intermediate phenotypes to link the genetics of addiction with that of AD. Such links might or might not have been anticipated, based on equivocal evidence from current epidemiologic studies.[150]

As we seek to document the extent of the generalization of effects of alleles that were initially identified in studies of addiction, we test the null hypothesis that clustered positive results from the GWA data from addiction vulnerability do not converge with the chromosomal positions of clustered nominally positive SNPs in comparisons of other phenotypes, such as individuals with bipolar disorder versus control samples. Monte Carlo simulations that test this null hypothesis sample from data within the SNP data sets noted above. As noted above, 100,000 trials allow estimates of the significance of the generalization of the effects of the alleles identified in studies of addiction vulnerability.

(e) ***Controls for the alternative possibilities that results could come from occult racial or ethnic stratification or assay noise.*** Several alternative hypotheses might explain observed results. To test some of these alternative hypotheses, we compare the clustered positive SNPs from different samples with SNPs that display the largest allele frequency differences in appropriate control data sets. These comparison data sets include those that contrast allele frequencies in (a) European American versus African American control individuals from NIDA samples,[12] (b) Japanese versus Han Chinese individuals from HapMap samples (JPT, Japanese from Tokyo; HCB, Han Chinese from Beijing), (c) control individuals sampled in different portions of the United Kingdom,[1] and (d) SNPs that display the largest variances from array to array.[12] We can thus compare data from the true comparisons in our experiments to similarly analyzed data from samples that test alternative hypotheses, providing substantial additional control evidence.

(f) ***Results from alternative approaches: principal components analysis and hierarchical clustering.*** Several alternative approaches to the analysis of GWA data sets can also provide interesting results that assess the structure of the pool-to-pool variance, based on data from 500 or 600K SNP sets. Principal components analysis (PCA) of the pool-to-pool differences in 500K data from European American, African American, and Asian samples divides the data along these racial and ethnic lines, as we might expect. However, PCA also subdivides data from experiments studying two distinct U.S. samples of nominally equivalent genetic background: Sample 1, NIDA European American subjects recruited in Baltimore, Maryland versus Sample 3, COGA European American subjects recruited in St. Louis, Missouri, the Bronx, New York, San Diego, California, and other sites. Similarly, these PCA analyses separate the samples of Asian subjects recruited in Japan from those recruited in or near Taipei who are self-characterized as Han Chinese. Each of these results underscores the need for extremely careful matching of the racial and ethnic backgrounds of control and disease samples.

Hierarchical clustering is most conveniently limited to data from individual genes. Large, 500K SNP data sets provide substantial limitations based on computer time. When we examine data from several genes using this approach in pools from a single racial or ethnic background, we can identify relatively clear patterns of separation between data from pools containing substance-dependent individuals and data from pools containing control individuals. These hierarchical clustering approaches are independent of the principal analyses noted above. These results reassure us that modest association signals can be

identified in many of these addiction-associated genes using a variety of different statistical approaches.

Ethical Issues in High-Density Genotyping of Individuals Selected Based on Self-Reported Illegal Behaviors

Individuals who are individually genotyped in relationship to addiction and related phenotypes are subject to a number of potential risks. Some of these risks are shared with individuals who are subjected to high-density genotyping in relationship to other disorders and phenotypes. Other risks are more likely to come to the fore in studies of illegal behaviors.

Concerns relating to insurability, employability, paternity determination, and providing (or not providing) genotyped individuals with access to their genotypes and/or genetic counseling are shared by individuals with other complex disorders.[151–153] Pending legislation in the United States may mitigate several of these concerns, and they are reviewed elsewhere. We therefore will not consider these issues further here.

High-density individual genotyping of DNA from individuals who are addicted to illegal substances raises additional issues. Many of these individuals are likely to have experienced involvement in criminal activities that goes beyond the use of illegal substances. Because the risks of high-density individual genotyping in this population have not been generally discussed elsewhere, we discuss several points that may inform thinking about these special ethical issues.

Increasingly ubiquitous DNA testing related to criminal activities lies at the heart of these concerns. In the United States, each state has a DNA database that collects information from crime scenes and from individuals convicted of particular offenses. A combined DNA index system (CODIS) operates local, state, and national DNA profile databases of convicted offenders, unsolved crime scenes, and missing persons. Numerous suspects have been identified through matches between DNA profiles from crime scenes and profiles from convicted offenders. A relevant website reports that the "success of CODIS is demonstrated by the thousands of matches that have linked serial cases to each other and cases that have been solved by matching crime scene evidence to known convicted offenders." The European Union is just one of the other international entities with a similar system (http://www.interpol.com/Public/Forensic/dna/dnafaq.asp).

Core CODIS data come from genotypes at 13 simple sequence length polymorphic (SSLP) loci. These loci lie near SNP markers with information about virtually all of these loci, providing a ready means of translating between SNP and SSLP genotypes. Other mitochondrial, sex chromosome, and autosomal markers are also genotyped on substantial numbers of these DNA samples.

A recent, October 2007, analysis of the CODIS-linked DNA index system revealed individually identifying genotype profiles for more than 5 million convicted offenders, as well as almost 200,000 DNA profiles from crime scenes (www.fbi.gov/hq/lab/codis/).

Almost 40% of men and 15% of women in cohorts from the areas of Baltimore from which Sample 1 and Sample 2 research volunteers come had experienced significant adult criminal justice system involvement (e.g., incarceration as adults) by the time they reached their late 20s (N. Ialongo, personal communication, 2008). It thus seems reasonable to conclude that several of the more than 3400 research volunteers from whom DNA was collected to form Samples 1 and 2 might be at potential risk for matches with crime scene DNA profiles. Similar potential risks might also be incurred through genetic study participation by individuals who report dependence on illegal substances in other parts of the United States. Although this problem is not unique to studies of the genetics of illegal behaviors, it appears to

be much more likely in this area than in most other areas of complex genetics.

Our laboratory, along with most other laboratories that work in this field, has established an elaborate means of coding, providing physical and electronic protections for the electronic and paper records that might identify our research volunteers. Subjects are protected by confidentiality certificates obtained through the Department of Health and Human Services. Data from these studies are analyzed and reported in ways that do not identify individual subjects.

However, the strongest protection for individuals who volunteer for this work comes from development and use of DNA pooling approaches. Because these approaches never generate high densities of genotypes for any individual, it is impossible to abuse or misuse these pooled data for unintended purposes. Pooling approaches provide these research volunteers with the strongest confidentiality protections currently available. Pooling may also merit increasing attention in other settings in which the risks of DNA-based personal identification are of significant concern.

Results: Clustering within Individual Samples and Convergence between Replicate Samples for the Same Phenotypes

Polysubstance Dependence versus Control (Samples 1 and 2)

NIDA substance dependence samples were analyzed[12] by selecting nominally positive SNPs that displayed P values < 0.05 for comparisons between substance-dependent and control samples within both European American and African American samples. We assessed the extent to which nominally positive SNPs that were identified in both samples by SNPs represented on at least two different array types cluster together in small chromosomal regions. Clusters contain at least three SNPs that display $P < 0.05$ in both samples and lie within 100 kb of each other. In this data set, 6666 of the 639,401 tested SNPs displayed reproducible, nominally significant abuser versus control allele frequency differences ($P < 0.05$) in both samples. The criterion that the same SNP display nominally significant association in each of two samples is more stringent than criteria used in other comparisons (see below). This criterion was applied to reduce the number of false positive results, but does not allow as much within-locus heterogeneity. Of these 6666 reproducibly positive SNPs, 1158 were within 320 chromosomal clusters; 184 of these clusters identified 244 annotated genes.

Monte Carlo simulation trials that assess the probability that these results are attributable to chance showed that none of the 100,000 simulation trials identified as many SNPs displaying nominally positive differences between substance-dependent and control samples in both European and African American samples as observed here (thus $P < 0.00001$).[12] Of the 100,000 Monte Carlo simulation trials, each of which began by selecting 6666 random SNPs, 2100 provided chromosomal clustering as marked as that observed for the true reproducibly positive SNPs ($P = 0.021$).

Methamphetamine-Dependent versus Control (Samples 4 and 5)

Nominally positive SNPs from each of these two samples clustered together, with no more than a 25-kb separation between nominally positive SNPs, more than anticipated by chance. In Sample 4, 846 clusters contained 3749 of the 15,569 nominally positive SNPs and, in Sample 5, 1787 clusters contained 8388 of the 25,538 nominally positive SNPs. Such clustering was not found in any Monte Carlo simulation trial ($P < 0.0001$ for both Sample 4 and Sample 5).

When we evaluated the genes identified by clustered, nominally positive results from Samples 4 and 5, we obtained evidence for

replication and results that could not be expected by chance alone. This criterion for genes to be identified by clustered, nominally positive SNPs from each of two samples is not as stringent as the criterion that the same SNPs produce nominally positive results in each of two samples. It does allow for within-locus heterogeneity. The degree of convergent identification of genes by data from each of these two samples was never observed by chance in any of 100,000 Monte Carlo simulation trials ($P < 0.00001$).

Bipolar Disorder versus Control (Samples 7, 8 and 9)

WTCCC Bipolar Disorder versus Control

Of the 426,604 SNPs analyzed in the WTCCC bipolar disorder collection, 28,192 displayed χ^2 values with $P < 0.05$.[1] Of these, 12,560 SNPs fell into 1775 clusters in which at least 4 SNPs, each of which displayed $P < 0.05$ (and were sampled on at least two array types), lay within 25 kb of each other. Monte Carlo simulation trials that assessed the probability that these results were attributable to chance showed that none of the 100,000 simulation trials identified as many clusters of SNPs that displayed nominally positive differences between bipolar disorder and control (control and other disease) samples as were actually identified (thus, $P < 0.00001$).[1]

NIMH Bipolar Disorder versus Control

Of the 536,288 SNPs analyzed in the NIMH bipolar collection,[2] 32,835 displayed t values with $P < 0.05$. Of these SNPs, 9971 fell into 1770 clusters in which at least 4 SNPs, each of which displayed t values corresponding to $P < 0.05$, lay within 25 kb of each other. Monte Carlo simulation trials that assessed the probabilities that these results were due to chance found that none of the 100,000 simulation trials identified as many clustered SNPs that displayed nominally positive differences between bipolar disorder and control samples as were actually identified from this work (thus, $P < 0.00001$).

German Bipolar Disorder versus Control

Of the 532,835 SNPs analyzed in the German samples, 27,057 displayed t values that corresponded to $P < 0.05$. Of these SNPs, 6110 fell into 1137 clusters in which at least 4 SNPs, each of which displayed t values corresponding to $P < 0.05$, lay within 25 kb of each other.[2] Monte Carlo simulation trials that assessed the probability that these results were due to chance found that none of the 100,000 simulation trials identified as many clustered SNPs that displayed nominally positive differences between bipolar disorder and control samples as were observed in this work (thus, $P < 0.00001$).

Convergent Data from at Least Two of the Three Bipolar Disorder versus Control Comparisons

Simulation trials assessed the likelihood that the clusters of nominally positive SNPs from at least two of these three bipolar samples identified the same genes. Monte Carlo simulation trials that assessed the probability that these results were attributable to chance found that none of the 100,000 simulation trials identified as many clustered SNPs displaying nominally positive differences between bipolar disorder and control samples in the same genes in multiple samples as we actually observed (thus, $P < 0.00001$).

Frontal Brain Volume (Samples 10 and 11)

In 500K Affymetrix data from unrelated members of NHLBI twin pairs, 10,266 SNPs provided nominally positive results (G.R. Uhl *et al.*, submitted). Of these SNPs, 583 fell into 169 chromosomal clusters, each of which contained at least three nominally positive SNPs that lie within 25 kb of each other and come from both the *Sty* I and *Nsp* I array types. Monte Carlo trials did not identify such a degree of clustering

in any of the 100,000 simulation trials (thus, $P < 0.00001$).

In 100K Affymetrix data from Framingham samples, 5786 SNPs provided nominally positive results.[8] Of these SNPs, 2729 fell into 613 chromosomal clusters with at least 3 SNPs within 100 kb of each other. Monte Carlo trials did not identify this degree of clustering in any of the 100,000 simulation trials (thus, $P < 0.00001$).

Only 212 of 100,000 Monte Carlo trials provided random fits between NHLBI and Framingham data sets that were as strong as those identified from the true data sets. Thus, Monte Carlo $P = 0.00212$ for the convergence between Sample 10 (500K) and Sample 11 (100K) data sets.

Success in Smoking Cessation (Samples 12, 13, 14)

In comparing data from successful versus unsuccessful quitters from Samples 12, 13, and 14, we identified 5411, 4539, and 4894 SNPs whose allele frequencies differed between these successful and unsuccessful quitters with nominal $P < 0.01$.[36] These nominally positive SNPs clustered together to extents much greater than expected by chance if their allele frequencies were independent of each other (Monte Carlo $P < 0.00001$). For Sample 12, 1434 of the 5411 nominally positive SNPs lay in 308 clusters in which each positive SNP was within 100 kb of at least two other positive SNPs with representation from SNPs on both array types. For Samples 13 and 14, 2258 of the 4539 nominally positive SNPs and 2184 of the 4894 nominally positive SNPs lay in 820 and 861 clusters, respectively, along with at least one other nominally positive SNP. We thus observed clustering in each of these independent samples. We would anticipate such clustering if many of these nominally positive SNPs identified haplotypes that were present in different frequencies in our samples of successful versus unsuccessful quitters, but not if they represented chance independent observations.

Convergent Data from at Least Two of the Three Smoking Cessation Success Comparisons

Nominally positive clustered SNPs from successful versus unsuccessful quitter comparisons from Samples 12–14 also clustered together on small chromosomal regions to extents much greater than expected by chance.[36] The Monte Carlo P values for the replication for samples 12 versus 13, 12 versus 14, and 13 versus 14 are 0.00054, 0.0016, and 0.00063, respectively.

Cognitive Ability (Samples 15 and 16)

Of the 489,922 autosomal SNPs assessed in Sample 15, 46,361 displayed correlations with cognitive ability assessments with nominal $P < 0.05$ (G.R. Uhl *et al.*, submitted). Of these nominally significant SNPs, 19,781 fell into 4016 clusters in which at least three nominally positive SNPs sampled on two array types lay within 25 kb of the neighboring nominally positive SNP. Similarly, of the 489,922 autosomal SNPs that were assessed in Sample 16, 36,689 displayed correlations with cognitive ability with nominal $P < 0.05$. Of these nominally significant SNPs, 11,376 fell into 2556 clusters. None of the 100,000 Monte Carlo simulation trials that sampled at random from the same databases found this degree of clustering of nominally positive SNPs by chance for either sample ($P < 0.00001$).

In data from each of the two cognitive ability samples, 348 genes contained at least three clustered nominally positive SNPs (G.R. Uhl *et al.*, submitted). Monte Carlo simulations never identified convergence of this magnitude by chance ($P < 0.00001$).

Alzheimer's Disease (Samples 17 and 18)

We found significant clustering of the SNPs that displayed nominally positive data in the GWA studies of AD.[4,5] For Sample 17, 5428

of 16,224 nominally positive SNPs identified 1125 clusters (reanalysis of 4). For Sample 18, 7469 of the 21,600 nominally positive SNPs fell into 1436 chromosomal clusters (reanalysis of 5). We never observed the extents of clustering identified here in any of the 100,000 Monte Carlo simulation trials for either sample ($P < 0.00001$).

Monte Carlo simulations also document the significant convergence between the results from these two independent studies of AD. Clustered nominally positive SNPs from both samples identified 99 genes (Uhl *et al.*, submitted). The observed overall convergence between these two data sets was never observed by chance in any of 10,000 simulation trials (thus, Monte Carlo $P < 0.0001$).

Generalization of Selected GWA Results between Related Disorders and Phenotypes

Convergence between Human GWA Studies of Dependence on a Variety of Addictive Substances (Samples 1–6)

We found a remarkable degree of overall convergence between the clustered, reproducibly positive SNPs in the NIDA samples and those identified by comparisons of alcohol-dependent and methamphetamine-dependent individuals[32] (Fig. 6; Tables 1 and 2).

The "methamphetamine dependence" genes described above displayed convergence with genes identified by (a) clustered positive results from GWA studies of polysubstance abuse in NIDA–MNB (Molecular Neurobiology) European American and African American samples,[12] (b) nominally positive SNPs from GWA studies of alcohol dependence,[3,33] and (c) nominally positive SNPs in comparisons of nicotine-dependent versus nondependent smokers.[3] Data from Samples 5 and 6 converged with these other addiction vulnerability data sets with Monte Carlo P values of 0.0412, 0.0016, and 0.0003, respectively.

Convergence between Human GWA Studies of Dependence on Addictive Substances and Success in Smoking Cessation (Samples 5 and 6 versus 12–14)

A number of the reproducibly positive genes identified by studies of methamphetamine dependence were also identified by clustered positive results from studies of Samples 12, 13, and 14 of European American smokers who were successful versus unsuccessful in abstaining from smoking during clinical trials for smoking cessation. P values were 0.002[34] for convergence with Sample 13 and < 0.00001 for convergence with Samples 12–14 taken together[36,37] (Fig. 5).

Bipolar Disorder (Samples 7, 8 and 9) versus Substance Dependence (Samples 1 and 2) Genes

We found remarkable convergence between the genes identified by clustered, nominally positive SNPs for addiction and those identified in a similar fashion for bipolar disorder (Uhl *et al.*, submitted; reanalysis of 1, 2). Few of the 100,000 Monte Carlo simulation trials identified chance overlaps as significant as those noted in actual data for comparisons between the genes identified by clustered positive SNPs for the NIDA substance dependence versus control comparisons and any of the three bipolar disorder versus control comparisons. Monte Carlo P values were < 0.00001 for the overlaps between data from Samples 1 and 2 versus data from Samples 7 and 9. The simulation P value was $P = 0.00012$ for comparison between data from Samples 1 and 2 versus Sample 8.

Neuroticism (Sample 19) versus Substance Dependence (Samples 1 and 2) Genes

We found significant convergence between the genes identified by clustered, nominally positive SNPs for addiction and those identified in a generally similar fashion for association

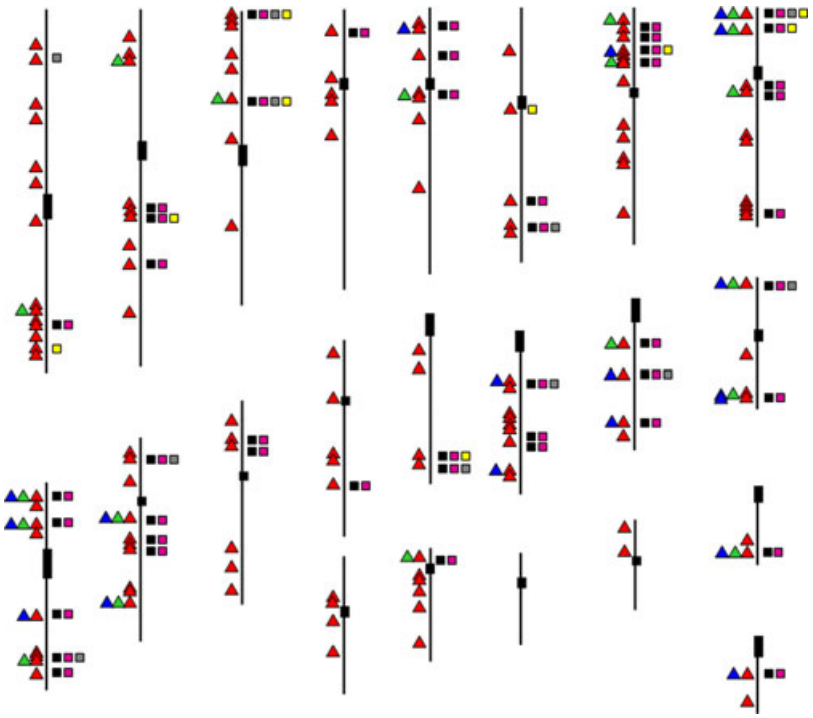

Figure 6. Distributions of data mapped onto cartoons of human chromosomes. Chromosomes 1–8 (*top row*), 9–16 (*second row*), 17–21 (*third row*), and 22 (*fourth row*). *Red triangles to the left of the (black) main axis* for each chromosome mark locations where genes are identified by clustered positive SNPs from Samples 1 and 2 and support is provided from at least one other sample. *Green triangles to the left of the main axis* mark locations where the gene is also identified by clustered positive SNPs from the methamphetamine samples (Samples 4 and 5). *Blue triangles to the left of the main axis* mark locations where the gene is also identified by clustered positive SNPs from two of the three nicotine abstinence samples (Samples 11–14). *Navy squares to the right of the main axis* mark locations where the gene is also identified by clustered positive SNPs from two of the three bipolar disorder samples (Samples 7–9). *Purple squares to the right of the main axis* mark locations where the gene is also identified by clustered positive SNPs from both cognitive function samples (Samples 15 and 16). *Grey squares to the right of the main axis* mark locations where the gene is also identified by clustered positive SNPs from both Alzheimer's disease samples (Samples 17 and 18). *Yellow squares to the right of the main axis* mark locations where the gene is also identified by clustered positive SNPs from the NHLBI frontal brain volume sample (Sample 10). Chromosomal positions are based on National Center for Biotechnology Information MAPVIEWER Build 36.1 coordinates and supplemental data from NETAFFX.

with neuroticism (Uhl *et al.*, in preparation, reanalysis of 121). Of the 31,110 SNPs with nominal $P < 0.05$ for comparisons of individuals with high versus low neuroticism scores, 4623 SNPs were within 903 clusters, each of which contained at least 4 such SNPs.

When we compared these clusters to those identified in Samples 1 and 2, only 30 of the 10,000 Monte Carlo simulation trials identified chance overlap that was at least as robust as the true data sets ($P = 0.003$).

Frontal Brain Volume (Samples 10 and 11) versus Cognitive Function (Samples 15 and 16) Genes

The frontal brain volume genes identified by both the NHLBI and Framingham data sets

TABLE 1. Cell Adhesion–Related Genes Identified in Multiple Genome-Wide Association Studies of Addiction and Related Disorders[a]

Gene	Description	Chr	bp	Samples with clustered positive SNPs	*P*
BAI3	brain-specific angiogenesis inhibitor 3	6	69404158	1,2,3,7,9,10,19	<0.00001
CDH13	cadherin 13	16	81218079	1,2,3,4,5,6,7,8,9,11,12,13,14,15,16,19	<0.00001
CLSTN2	calsyntenin 2	3	141136897	3,4,5,7,8,10,12,13,14,15,16,17,18,19	<0.00001
CNTNAP2	contactin-associated protein-like 2	7	145444386	3,4,5,6,7,8,9,11,14,15,16,18,19	<0.00001
CSMD1	CUB and Sushi multiple domains 1	8	2782789	1,2,3,4,5,6,7,8,9,10,11,13,14,15,16,17,18,19	<0.00001
CTNNA2	catenin α 2	2	79593634	3,4,5,6,7,8,9,13,14,15,16,19	<0.00001
DAB1	disabled homolog 1	1	57236167	3,4,5,6,7,8,9,12,14,15,16,18,19	<0.00001
DSCAM	Down syndrome cell adhesion molecule	21	40306213	1,2,3,4,5,7,8,12,13,15,16,17,19	<0.00001
NRXN1	neurexin 1	2	50000992	3,4,5,6,9,11,14,15,16,17,18,19	<0.00001
PTPRD	receptor protein tyrosine phosphatase D	9	8307268	1,2,3,4,5,6,8,11,13,14,15,16,17	<0.00001
SGCZ	sarcoglycan zeta	8	13991744	1,2,3,4,5,8,9,10,11,13,14,15,16,17,19	<0.00001
ASTN2	astrotactin 2	9	118227328	1,2,4,5,6,8,9,13,15,16,17,18,19	0.000070
CNTN4	contactin 4	3	2117247	1,2,3,6,8,9,11,15,16,17,18	0.000110
CNTN6	contactin 6	3	1109629	1,2,3,4,6,8,9,10,15,18	0.000120
LRP1B	low-density lipoprotein-related protein 1B	2	140705466	1,2,3,5,6,8,9,10,11,13,15,16,18,19	0.000150
NRG1	neuregulin 1	8	32525295	3,4,5,6,7,11,15,16	0.000240
ITGB8	integrin β 8	7	20337250	1,2,8,9,15,16	0.000260
PTPRM	receptor protein tyrosine phosphatase M	18	7557817	1,2,3,4,5,7,11,13,15,16,18,19	0.000290
ROR1	receptor tyrosine kinase-like orphan rec 1	1	64012302	4,5,7,8,9,13,15,16,18	0.000290
TRIO	triple functional domain/PTPRF interact	5	14196829	1,2,9,11,13,14,15	0.000690
CSMD2	CUB and Sushi multiple domains 2	1	33752196	1,2,3,4,7,8,11,14,17,18,19	0.000830
CNTN5	contactin 5	11	98397081	1,2,3,5,6,7,8,9,11,14,15,19	0.000980
CTNNA3	catenin α 3	10	67349937	1,2,3,5,6,7,8,13,15,16,19	0.001090
LRRN6C	leucine rich repeat neuronal 6C	9	27938528	1,2,3,4,5,6,7,8,9,11,15,16,17,19	0.001340
CTNND2	catenin δ 2	5	11024952	1,2,3,5,6,9,15,16	0.003270
ANKS1B	ankyrin repeat sterile α domain 1B	12	97653202	1,2,5,6,9,15,16,19	0.003410
SEMA3C	semaphorin 3C	7	80209790	1,2,9,16	0.006310

[a]Columns list gene symbol, gene description, chromosome, basepair of gene start, samples (from this review) in which clustered nominally positive SNPs are found, and overall *P* value for this gene in this entire data set, based on 100,000 Monte Carlo simulation trials. Note that genes that are identified strongly in a single sample will not be included on this list.

TABLE 2. "Druggable" Genes Identified in Multiple Genome-Wide Association Studies of Addiction-Related Disorders[a]

Gene	Class	Description	Chr	Start	Samples	*P*
RYR3	CHA	ryanodine receptor 3	15	31,390,469	1,2,3,4,5,6,7,8,9,13,15,16,17,19	<0.00001
CAMK1D	ENZ	calcium/calmodulin-dependent protein kinase ID	10	12,431,589	1,2,7,8,9,13,15,16,17,18,19	<0.00001
CHN2	ENZ	chimerin 2	7	29,200,646	1,2,3,7,8,10,13,14,15,16	<0.00001
FHIT	ENZ	fragile histidine triad gene	3	59,710,076	1,2,3,4,5,6,7,8,9,10,13,15,16,17,18,19	<0.00001
PRKG1	ENZ	cGMP-dependent protein kinase I	10	52,504,299	1,2,3,4,5,6,7,8,9,11,12,13,14,15,16,18,19	<0.00001
SERPINA1	ENZ	serpin peptidase inhibitor A 1	14	93,914,451	1,2,5,13,14,18	<0.00001
LARGE	ENZ	like-glycosyltransferase	22	31,999,063	3,4,5,7,9,13,14,15,16,17,18	0.00002
UST	ENZ	uronyl-2-sulfotransferase	6	149,110,157	1,2,6,15,16,17,18	0.00017
DAPK1	ENZ	death-associated protein kinase 1	9	89,302,576	1,2,3,5,13,14,15,16	0.00024
GALNTL4	ENZ	UDPNAc-α-D-galactosamine:polypeptide NAcgalactosaminyltransferase-like 4	11	11,248,999	1,2,3,4,5,,9,15,16,19	0.00040
CAPN13	ENZ	calpain 13	2	30,799,141	1,2,3,5,6,8,15	0.00042
PDE4D	ENZ	cAMD-specific phosphodiesterase 4D	5	58,302,468	1,2,4,5,6,8,9,15,16,19	0.00055
AKAP13	ENZ	protein kinase A anchor protein 13	15	83,724,875	1,2,4,6,8,13,14,15,16	0.00062
ST6GALNAC3	ENZ	αNAcneuraminyl-2,3-β-galactosyl-1,3-NAcgalactosaminide α2,6-sialyltransferase 3	1	76,312,992	4,5,6,9,11,13,14,15,16,17,19	0.00063
SERPINA2	ENZ	serpin peptidase inhibitor A 2	14	93,900,404	1,2,13,14	0.00080
DGKB	ENZ	diacylglycerol kinase β	7	14,153,770	1,2,3,8,9,13,15,16,19	0.00138
FGF14	LIGAND	fibroblast growth factor 14	13	101,173,036	1,2,3,6,7,12,15,16,17,18	0.00001
GRM7	REC	metabotropic glutamate receptor 7	3	6,877,927	1,2,3,5,6,7,8,11,14,15,17	<0.00001
GPR39	REC	G protein–coupled receptor 39	2	132,891,086	1,2,8,14,15,18	0.00033
ABCC4	TRANSP	ATP-binding cassette C 4	13	94,470,090	1,2,5,10,15,16,17	<0.00001
SLC24A3	TRANSP	solute carrier family 24 member 3	20	19,141,290	3,4,5,7,8,11,14,15,16,18	0.00033
XKR4	TRANSP	XK family member 4	8	56,177,571	1,2,3,4,5,11,15,16,19	0.00091
SLC9A9	TRANSP	solute carrier family 9 member 9	3	144,466,754	1,2,3,4,6,8,9,13,15,18	0.00134

[a]Columns list gene symbol, gene description, chromosome, basepair of gene start, samples (from this review) in which clustered nominally positive SNPs are found and overall *P* value for this gene in this entire data set, based on 100,000 Monte Carlo simulation trials. Note that genes that are identified strongly in a single sample will not be included on this list.

overlapped with those identified in GWA studies for cognitive function, the genetics of which is linked to frontal lobe function (Uhl *et al.*, submitted).

Nineteen genes identified by clustered nominally positive SNPs for frontal lobe volume in the NHLBI data set were also identified by clustered nominally positive SNPs from each of the two GWA assessments for cognitive function. Such overlap was identified by chance in only 212 of 100,000 Monte Carlo simulation trials that correct for repeated comparisons ($P < 0.00212$). One of these genes, CLSTN2 was also identified in a GWA study of individual differences in memory function.[154]

Alzheimer's Disease (Samples 17 and 18) versus Cognitive Function (Samples 15 and 16) Genes

Each of these overlapping AD data sets overlapped significantly with each of the data sets for cognitive function. None of the 100,000 Monte Carlo simulation trials identified as many overlapping genes as those identified by the true data sets (Monte Carlo $P < 0.00001$) (Uhl *et al.*, submitted). These comparisons thus strongly support generalization of the observations that we have made in studying general cognitive ability to vulnerability to AD. Pleiotropic effects of variants in a number of genes are thus likely to influence both cognitive ability and AD.

Alzheimer's Disease (Samples 17 and 18) versus Frontal Brain Volume (Sample 10) Genes

These overlapping AD data sets overlapped significantly with the data set for frontal brain volume (Uhl *et al.*, submitted). None of the 100,000 Monte Carlo simulation trials identified as many overlapping genes as those identified by the true data sets (Monte Carlo $P < 0.00001$). Pleiotropic effects of variants in a number of genes are thus likely to influence both AD and frontal brain volume.

Alzheimer's Disease (Samples 17 and 18) versus Substance Dependence (Samples 1 and 2) Genes

Data from the overlapping AD data sets also overlapped significantly with data from the overlapping NIDA addiction vulnerability samples (Samples 1 and 2).[12] None of the 100,000 Monte Carlo simulation trials identified as many overlapping genes as those identified by the true data sets (Monte Carlo $P < 0.00001$). These comparisons thus strongly support the notion of transitive genetic overlaps between addiction, cognitive function or brain volume, and AD based on pleiotropic effects of variants in a number of genes.

Failure of Control Experiments to Support Alternative Hypotheses for the Observed GWA Results

We found no evidence that many of the clustered, reproducibly positive SNPs identified in these sample sets resulted from alternative effects of several different classes.

We found no significant overlaps between the SNPs identified in these studies and the SNPs that provided the largest racial and ethnic differences. Comparison groups included (a) SNPs that provided the largest racial and ethnic differences in comparisons between European American and African American control individuals from Samples 1 and 2; (b) for Asian samples, SNPs that provided the largest racial and ethnic differences between Japanese and Chinese HapMap samples; and (c) for samples of European heritage, SNPs that displayed the largest allele frequency differences between U.K. samples of self-reported European heritage.[155]

We found no significant overlaps between the SNPs identified in these studies and the SNPs that displayed the largest variances from array to array. Thus, we found no significant support for either of these alternative hypotheses: (a) that the positive results of any of these studies resulted from occult racial and ethnic

stratification or (b) that the positive results of these studies resulted from assay noise alone.

Finally, comparisons with data from a complex phenotype outside the brain provide a useful contrast with the observed convergence identified in the data sets above. Data for type I diabetes from the WTCCC provide such a contrast.[1] Using Monte Carlo simulation approaches identical to those used to identify the highly significant overlaps noted above, we identified no significant overlap between genes that contain clustered, reproducibly nominally positive SNPs for substance dependence (Samples 1 and 2) and genes that have been identified in a similar fashion for type I diabetes (C. Johnson and G. R. Uhl, unpublished observations, 2007).

Limitations and Cautions

It is important to consider several limitations for these convergent, replicated genome-wide data. The sizes available for many of the samples from which data are reviewed here provide moderate power, at best, to detect gene variants. False negative results are likely because we require positive data from each of the several samples. The likelihood of false negatives is also increased because we require positive results from several SNPs from at least two array types that cluster within small chromosomal regions, making it easier to miss modest association signals within small genes that contain few SNPs or genes whose SNPs lie on only one array type. We have recently reassessed the relatively large numbers (6734) of currently annotated genes for which there is no possibility of detection, using 500K array data and the criterion that a gene must be identified by "at least three clustered (25 kb) SNPs from at least two array types lying within the gene's exons, introns or 10 kb flanking sequences." Genes that do not have the requisite numbers of SNPs with these characteristics are listed in the Supplement in Table SII. This list includes genes for which substantial positive association data are available from candidate gene studies, including the DRD4 dopamine receptor.[156–158] A number of these genes could be detected with the addition of 100 and 500K data, and even more could be detected with more recently available arrays that assess approximately 1 million SNPs; these genes are also indicated in this table.

Differences in allele frequencies in different populations could explain why some genes are strongly associated with substance dependence in European but not in African or Asian samples. Apparent failure to replicate in the present analyses could thus derive from the differential informativeness of SNP markers in samples from different racial or ethnic backgrounds.

We focus only on data from autosomal regions. This focus allows us to combine data from male and female subjects. However, such analyses will miss potentially important contributions from genes on sex chromosomes.

Many of the subjects of these studies volunteered for demanding clinical protocols, potentially rendering them not totally representative of all individuals in the population with the same phenotypes. Some of the subjects for this work have also come to clinical attention because of specific clinical features, including presentation to emergency rooms with methamphetamine psychosis or presentation to memory clinics with memory complaints. These selection criteria might not render these individuals completely representative all individuals who display the phenotype or disorder.

Ethical considerations arise in studying substance abuse, an illegal behavior often associated with other difficult individual and family pathologies (see also above). Such ethical considerations suggest that it is unlikely that any large, truly population-based sample for dependence on illegal substances will be readily obtained. Sampling research volunteers and examining the extent to which the volunteers' demographic characteristics match or deviate from those of the larger drug abuser communities provides one approach to this issue. We cannot quantitate the selection biases that might be

imposed by factors not reflected in the demographic features (e.g., gender and self-reported racial or ethnic background) that are often used to characterize subgroups.

For studies of the genetics of brain volume, it is worth noting that at least some of the genetic influences on frontal cortical volume are likely to provide pleiotropic influences on other brain regions. Some gene variants that influence frontal lobes as well as addiction, AD, and cognitive functions could even have effects on these behavioral phenotypes through actions outside of the frontal cortex.

For the current analyses, we employed principal analytic approaches that represent only one of many current approaches to analyzing GWA data (see discussion above, also at www.nhlbi.nih.gov/resources/listserv/number 35.htm [RFA-HL-07-010]).[6,85] Despite this limitation, the replicated positive results obtained here and the failure of control experiments to support alternative hypotheses do provide significant confidence in the roles of most of the genes reported.

The generalization of results from replicated samples to data from other samples that are likely, *a priori*, to display genetic overlap also increases the confidence in these results. Although it is not possible to supply precise estimates for this enhanced confidence, the generalization reported here does add to the overall confidence in the sets of genes listed. Identifying the same gene in multiple studies also enhances confidence that the gene contains relatively common allelic variants that alter its function sufficiently to produce association signals in several GWA data sets.

However, it is important to note again that, although the analytic approach that we use here seeks to identify genes with variants that are likely to play roles in addiction and in related phenotypes, we believe that the interpretation of these data is limited. Flexible approaches to the identification of such genes are likely to be useful in analyses such as those that are presented here. If we allow for locus heterogeneity, for sample-to-sample differences in the details (including the phase) of linkage disequilibrium between marker alleles and pathogenic alleles, and for the use of different marker sets with different data sets, then we cannot demand that all samples display the more stringent criterion—a significant association of the same SNP with the same phase in all samples. This approach thus seeks to use stepwise analyses to identify and confirm a statistically significant association at the level of individual gene loci, assuming that subsequent follow-up analyses will use further steps to untangle any locus heterogeneity, to unequivocally identify which individual SNPs are associated, and to identify pathological haplotypes and the phases with which they are associated with phenotypes in samples from different racial and ethnic backgrounds. Indeed, we believe that, in many circumstances, molecular biologic, behavioral, and other evidence will be required to have confidence in pathological haplotypes; in this way, the analyses that we present here represent open invitations for just these sorts of ancillary studies to guide further molecular genetic fine mapping efforts.

For diagnoses of individuals selected because of their dependence on a specific substance, such as methamphetamine, the specificity of effects of a gene's variants is probably limited by the fact that many of the subjects for these studies also report use of additional addictive substances (e.g., inhalants for the methamphetamine dependence Samples 4 and 5). These clinical considerations, as well as the overlap between the methamphetamine dependence genes and the genes identified in other GWA work, support the idea that many, but not all, of the methamphetamine dependence (or nicotine dependence) loci probably contain allelic variants that provide a more general vulnerability to addictive substances. We refer to some genes as *methamphetamine dependence genes*, for example, to denote the fact that variants in these genes are likely to alter vulnerability to developing dependence on this substance; however, many of these allelic variants may also predispose individuals to dependence on other addictive substances.

Relatively few of the control subjects in the studies of substance dependence report any significant use of illegal addictive substances. The genes identified herein thus could influence vulnerabilities to a number of features important for eventually developing dependence, including initiation of use, persistence of use, failure of attempts to quit, transition from persistent use to dependence, or other steps, in ways that are not completely unraveled by simple comparisons of dependent versus control samples.

The largest limitation to the analyses presented here, however, probably comes from the modest size of the effects reported for each of these phenotypes (with the exception of the APOE influences in AD, of course). In each polygenic disorder, replication of the effects of polygenic loci is unlikely to be sufficiently robust that each study will provide nominally significant observations. Final confidence in true effects requires both replication and generalization, including studies of additional samples with the phenotypes that we considered in the present review. However, we also need to display caution in considering the significance of apparent nonreplications. Recent analyses underscore the role that heterogeneity can play in engendering false negative results from attempts at replication.[126]

Classes of Genes That Are Identified in Multiple GWA Samples for Multiple Phenotypes: Focus on Cell Adhesion–Related Genes

One approach to describing the convergence between the data sets presented above relies on the overall convergence between the results obtained in each study. A different approach focuses on convergence of data concerning specific genes and classes of genes, especially when most are expressed in the brain. Many of the genes that we identify in this analysis of convergent GWA findings are involved in cell adhesion processes whereby neurons recognize and respond to features of their environments that are important for establishing and maintaining proper connections (Table 1). Others are involved in enzymatic activities; protein translation, trafficking, and degradation; transcriptional regulation; receptor, ion channel, and transport processes; disease processes; and cell structures. A subset of this latter group is of special interest because it represents a group of classically "druggable" targets for potential small-molecule therapeutics (Table 2).

Cell Adhesion–Related Genes

The genes whose products are involved in cell adhesion processes provide a number of especially interesting results (Table 1). Cell adhesion mechanisms are central for properly establishing and regulating neuronal connections during development. Cell adhesion mechanisms can play major roles in mnemonic and other neuroadaptive processes in adults.[159,160] Interestingly, most of the cell adhesion–related genes that we identify in these GWA studies are expressed in developing and adult brains. Altered expression of several of these genes can alter neurite extension,[161–163] activate signaling pathways,[164–169] and alter mnemonic processes.[163] Almost all of these cell adhesion–related genes are expressed in memory-associated brain regions that include the hippocampus and cerebral cortex (http://brain-map.org).[170–173] By contrast, substantial expression in mesolimbic/ mesocortical dopamine reward system neurons is not documented for many of these genes.

Cell adhesion–related genes identified by these GWA studies encode members of several structural cell adhesion molecule (CAM) subfamilies, including those that are anchored to cell membranes by glycophosphoinositol (GPI) anchors, those that display apparent single-transmembrane topologies, those that display apparent seven-transmembrane topologies, and those that produce soluble products.

Here, we discuss several of the cell adhesion–related genes that have been identified in multiple sets of replicated GWA studies (Table 1).

Monte Carlo trials that seek to establish the level of significance of identification of these genes in multiple data sets establish levels for this convergent identification that range from 0.006 to < 0.00001. We also discuss follow-up data for two CAM genes, NrCAM and NRXN3, which were initially identified by lower-density GWA studies.[10,11] Results from these two loci provide examples of the ways in which initial identification in GWA studies can lead to subsequent identification of specific haplotypes that alter levels of expression of the gene or levels of expression of a specific set of splice variants.

CAMs with the Strongest Levels of Cumulative Support

One of the CAMs that achieves the most striking nominal *P* values in these analyses is an atypical member of the cadherin gene family, Cadherin 13 (CDH13). A GPI-anchored CAM, CDH13 is expressed in neurons within brain regions that are likely to play roles in addiction, including the hippocampus, frontal cortex, and ventral midbrain.[171] Cadherin 13 can inhibit neurite extension from select neuron populations[161,171] and activate a number of signaling pathways.[164–167] It is thus a strong candidate for roles in brain mechanisms that are important for both developing and quitting addictions.

Other cell adhesion–related genes that manifest cumulative *P* values less than 10^{-6} in these analyses include BAI3, CLSTN2, CNTNAP2, CSMD1, CTNNA2, DAB1, DSCAM, NRXN1, PTPRD, and SGCZ. We discuss several of these genes here.

DSCAM is a single-transmembrane domain CAM with immunoglobulin and fibronectin domains that is expressed strongly in the brain,[169,174] and specifically in the hippocampus, in ways that are required for appropriate neuronal connections to form in memory-associated circuits in model organisms.[162,163] Different dendritic processes of the same neuron do not often cross each other; this self-avoidance mechanism depends on the expression of a large array of tightly regulated DSCAM isoforms.[175,176] Simplifying this repertoire substantially disrupts appropriate formation of neuronal networks *in vivo*.[177] Indeed, flies with altered DSCAM expression display altered memories for both rewarded and punished behaviors.[163]

CLSTN2 contains allelic variants that are identified in GWA studies of individual differences in memory and executive function as well as in the cognitive ability, AD vulnerability, and frontal brain volume phenotypes reviewed here.[178,179] CLSTN2 is expressed in the frontal cortex and hippocampus[172] and is well-positioned to provide calcium-dependent cell adhesion functions in brain regions that include the hippocampus and in the postsynaptic densities where it is highly expressed. The structure and expression of CLSTN2 make it a good candidate to function as a single-transmembrane domain CAM, variants of which could alter the ways in which neuronal and synaptic connections develop, the ways in which they are maintained and reorganized in adult brains, or both.

DAB1 interacts with and participates in signaling from several CAMs; it has long been identified with signaling through the CAM reelin in ways that alter formation and maintenance of neuronal processes.[180] More recent evidence also supports roles for DAB1 in signaling through other cell adhesion and cell regulatory mechanisms, including those that use the amyloid precursor protein CAM.[181] DAB1 is expressed in many brain neurons, including those in the hippocampus and in mid to deep cerebral cortical layers[172] (http://brainmap.org). Mice with DAB1 disruption display substantial alterations in cerebral cortical development accompanied by gross motor and other behavioral phenotypes.[182]

BAI3, a seven-transmembrane domain CAM, and PTPRM, a single-transmembrane receptor tyrosine kinase that mediates homophilic cell recognition and is supported at a more modest level of statistical confidence,

are both expressed in vasculature.[183,184] Identification of these genes fits with the idea that control and regulation of angiogenesis and vascular functions play important roles in determining the richness of the cerebral cortex and other areas of adult brains[184] in ways that have consequences for a variety of interesting brain-based phenotypes. In addition, substantial neuronal expression of PTPRM occurs in cortical and cerebellar cortical neurons.[172,183]

PTPRD is expressed in the brain and displays prominent hippocampal expression. Its extracellular ligands have not been elucidated, although it can bind to liprin.[185] PTPRD-knockout mice display altered hippocampal long-term potentiation and spatial learning;[186] these findings fit well with the human phenotypes related to cognitive function. Mice with deletions of both PTPRD and a related PRP sigma (but not with either knockout alone) die at birth because of a failure to innervate appropriately.[187] SGCZ participates in protein complexes with cell adhesion–like.[188] High levels of SGCZ expression in the brain are confirmed by Allen brain atlas images.[172] Biochemical studies identify expression in the Schwann cells of peripheral nerves.[188] SGCZ can be found in complexes with αδ or with εβδ sarcoglycans, demonstrating specificity of the context of its function in brain.[189]

CSMD1 is substantially expressed in adult brain regions that include the hippocampus.[170] High levels of CSMD1 expression in the growth cones of neurons cultured from the developing brain support substantial roles in development as well.[170] Less striking levels of evidence implicate variants in CSMD family members CSMD2 and CSMD3 in several of these brain-related phenotypes.[190]

CAMs with Intermediate Levels of Cumulative Support

Cell adhesion–related genes with nominal cumulative significance levels between 0.00007 and 0.006 include ANKS1B, ASTN2, CNTN4, CNTN5, CNTN6, CTNNA3, CTNND2, LRP1B, LRRN6C, NRG1, ITGB8, PTPRM, ROR1, TRIO, CSMD2, and SEMA3C. We discuss several of these molecules here.

ASTN family members play substantial roles in the migration and organization of neuronal processes along glial fibers in ways that are key to proper development of laminated brain structures.[191] ASTNs are expressed at high levels in the adult hippocampus.[172] ASTN knockouts exhibit slower development of these structures.[192] ASTNs are upregulated after hippocampal lesions that induce synaptogenesis.[193]

CNTN6 is a GPI-anchored CAM that can act as a Notch ligand, triggering nuclear translocation of the Notch intracellular domain. This interaction can promote formation of oligodendroglia from progenitor cells and can increase the expression of myelin-associated glycoprotein in oligodendrocytes.[194] Neuronal expression in hippocampal, cerebral cortical, cerebellar, and thalamic subdivisions in the adult brain could contribute to the motor incoordination phenotypes noted in knockout mice.[195,196]

SEMA3C is one of the CAMs that has been implicated in the development of dopaminergic projections.[197] SEMA3C and its neuropilin 2 receptor are regulated by neuronal injury.[198] Significant levels of SEMA3C expression in the hippocampus and connected regions, such as the medial septal nucleus, provide a good rationale for involvement in memory-related phenotypes.[199]

LRRN6C displays interesting patterns of brain expression, with dense expression in the cells of the dentate gyrus of the hippocampus and in entorhinal cortical regions.[172] However, little additional evidence concerning this gene has appeared in the literature to date.

CTNNA2 is another protein whose strong association with cell adhesion mechanisms, especially cadherin mechanisms, renders it of substantial interest here. Mice with impaired CTNNA2 expression display impairments in prepulse inhibition and fear conditioning,

altered dendritic spine morphogenesis in hippocampal neurons, unstable synaptic junctions, and defective anterior commissure formation.[200,201]

Potential Roles for Cell Adhesion–Related Genes

The cell adhesion genes identified here provide an attractive way to bridge the gap between (a) the remarkable observed overlap between the molecular genetics of the clinical and cognitive phenotypes reviewed here and (b) brain differences, especially those that might manifest in the quantity and/or quality of neuronal connections, that might underlie these shared heritable influences.

These cell adhesion–related genes also provide an attractive bridge between the genetics of phenotypes that are not accompanied by gross brain pathology and those that are. Some of the cell adhesion genes that we identify in studies of cognitive ability might conceivably be directly involved in AD pathological processes.[202] However, the majority of the cognitive ability genes identified here fit with emerging views that relatively subtle differences in brain connections that contribute to individual differences in cognitive abilities also provide individual differences in *cognitive reserves*.[203] If cognitive reserves mitigate the cognitive impact from a given density of AD senile plaques and neurofibrillary tangles, for example, individuals with greater cognitive reserves might die without dementia despite neuropathological brain burdens that would otherwise produce significant dementia in individuals with lesser cognitive reserves.[204]

Cell Adhesion Molecule Genes with Substantial Follow-up Information

Neurexin 3. Neurexins are CAMs that help to specify and stabilize synapses and provide receptors for neuroligins, neurexophilins, dystroglycans, and α-latrotoxins.[168,205,206] Neurexins function in the nervous system as CAMs at excitatory and inhibitory synapses.[168,205–214] The mammalian neurexin genes, *NRXN1–NRXN3*, each display multiple promoters from which longer α-neurexins and shorter β-neurexins are transcribed. Differential promoter use and/or differential splicing events provide many neurexin isoforms.[213,215–218] Neurexin splicing variants also provide opportunities to produce both membrane-bound and soluble isoforms.[215–217]

NRXN3 expression in specific cerebral cortical regions and layers, especially layers 2–3 and 5–6, is documented in rodent brain atlas images.[172] Although *NRXN3* is also expressed in other brain regions of interest for addiction, its expression in glutamatergic projections that arise from the prefrontal cortex to innervate the nucleus accumbens and other striatal regions and/or in γ-aminobutyric acid (GABA)-ergic neurons that project from cortical area to cortical area provide sites at which altered expression could readily influence circuits that are key for addictive behaviors. Alterations in the size or strength of synapses in these circuits could readily provide alteration in behavioral features, including vulnerability to addictions.

From 10,000-SNP GWA data, we initially reported that SNP rs760288, located in the 3′ region of *NRXN3*, displayed allele frequencies that distinguished individuals dependent on illegal substances from control individuals in both Sample 1 and Sample 2.[11] These observations have received support from the results derived from Sample 6[3] and a linkage study of opioid dependence.[219]

We genotyped nine SNPs in the 3′ *NRXN3* regions that lie near two splicing sites, termed SS#4 and SS#5, in 144 European American alcohol-dependent individuals and 188 control individuals from Sample 3.[130] Minor allele frequencies at each of four SNPs near SS#5 were higher in alcohol-dependent than in control samples, supporting initial observations in other samples. A "T" allele at the rs8019381 (close to the original rs760288 SNP) produced a 2.46 odds ratio for belonging to the alcohol-dependent group. Common 3 SNP haplotypes in this area were also strongly associated with alcohol dependence (nominal $P = 0.000598$,

corrected to $P = 0.00580$ based on permutations).

The rs8019381 SNP was located 23 bp 3′ from the 3′ end of exon 23, near a branch point that could change the splicing of *NRXN3* primary transcripts.[220,221] In mRNAs isolated from postmortem cerebral cortex, individuals with one or two copies of the minor rs8019381 T allele expressed less of the major transmembrane isoform ex22a24b and the minor transmembrane isoform ex22a24a mRNA levels than did CC homozygotes ($P = 0.0008$ and 0.021 for these genotypes by a two-tailed Mann–Whitney test, respectively). By contrast, expression of the major ex22a23a soluble isoform did not correlate with the rs8019381 genotype ($P = 0.171$ by two-tailed Mann–Whitney test). As a consequence, ratios between transmembrane and soluble isoforms were significantly increased in individuals with one or two T alleles compared with CC homozygotes ($P = 0.018$, two-tailed Mann–Whitney test).

Standard models for the development and alteration of excitatory and inhibitory synapses posit that neurexins provide presynaptic CAMs that bind heterophilically and with calcium dependence to postsynaptic CAMs to form transsynaptic complexes that help to stabilize excitatory or inhibitory synapses.[205] Repertoires of diffusible entities enrich this picture. Neurexophilins provide soluble neurexin ligands that are expressed in neurons in a number of brain regions that also express *NRXN3*. Diffusible *NRXN3* isoforms are expressed in brain regions that express several neuroligins and dystroglycan.[222] Conceivably, soluble neurexins and neurexophilins might diffuse to sites at which membrane-bound neuroligins and neurexins are localized. Enhancing ratios between diffusible and membrane-bound *NRXN3* isoforms in the CT/TT genotypes of rs8019381 might thus provide more diffused signaling in ways that could alter pruning, spreading, and/or strength of synaptic contacts.[223,224]

NrCAM. *NrCAM* encodes a single-transmembrane domain CAM with six immunoglobulin domains, four to five fibronectin III repeats, a transmembrane domain and a C-terminal cytoplasmic domain with tyrosine kinase phosphoacceptor sites.[225] Multiple *NrCAM* isoforms can be identified as the products of differential RNA splicing events that produce *NrCAM* translation products that contain short inserted peptide sequences. *NrCAM* mRNA expression is high in the hippocampus, middle to deep layers of the cerebral cortex, cerebellar Purkinje cell layers, and striatal interneurons. Cells of the substantia nigra and ventral tegmental area (VTA) express hybridization densities that are almost as dense as those identified in hippocampal pyramidal neurons.[131] Many, but not all of the *NrCAM* immunopositive neurons in the VTA also express dopamine transporter (DAT) immunoreactivity, a marker for dopaminergic neurons.[131]

We initially identified NrCAM based on convergence between data from GWA and studies of drug-regulated gene expression.[226] Our initial 1.5 k SNP GWA study identified a reproducibly nominally positive SNP that mapped roughly to the region that contained NrCAM and other genes. Follow-up studies provided confirmatory observations: 11 of 37 tested mid chromosome 7 SSLPs displayed allele frequency differences between abusers and controls in NIDA European American and/or African American samples that reached nominal significance. Three displayed nominally significant differences in both populations; fewer than one would have been expected by chance. These nominally significant results encompassed markers located between 98 and 103 Mb of chromosome 7.

Subtracted differential display PCR (SDD) identified NrCAM as a morphine-regulated gene. Sequences that encoded 3′ untranslated regions of NrCAM were among the subcloned SDD cDNAs that corresponded to mRNAs whose expression was altered in the striata of rats sacrificed 4 h after treatment with 20 mg/kg morphine. These biochemical results led us to focus on NrCAM. We thus identified 3′ and 5′ blocks of restricted haplotype diversity

in NrCAM in European American and African American samples and sought association between addiction and allelic frequencies of markers in both of these haplotype blocks. The 3′ haplotype was associated with addiction vulnerability in European American individuals from Samples 1 and 3 (nominal $P = 0.0006$ and $P = 0.003$, respectively) and from African American individuals from Sample 2 (nominal $P = 0.0006$). However, the phase of the association was opposite in the African American samples compared with the phase in the two European American samples. Frequencies of the addiction-associated 3′ *NrCAM* haplotype were thus higher in African American abusers than in controls but lower in two samples of European American substance abusers than in the corresponding control samples. We interpreted these phase differences as indicating that the 3′ *NrCAM* haplotype block was close to, but not identical with, the pathogenic haplotype.

NrCAM 5′ flanking region haplotypes, however, displayed associations with the same phase in each of these three samples ($P = 0.0002$, $P = 0.06$, and $P = 0.02$, respectively). Further, an additional independent sample that compared haplotype frequencies in 288 alcohol-dependent Japanese subjects versus 472 matched controls also displayed a highly significant association ($P < 0.01$) with the same phase. These cumulative observations strongly support the idea that *NrCAM* variants provide polygenic contributions to human interindividual differences in addiction vulnerability. They contrast with the failure to identify such reproducibly positive findings at the adjacent genes, including LAMB1, LAMB4, and iPLA2.

We sought possible functional effects of these 5′ *NrCAM* haplotypes by assessing patterns of allele-specific expression of *NrCAM* haplotypes in mRNA extracted from human postmortem brains. The mRNA corresponding to the addiction-associated 5′ *NrCAM* haplotype was expressed at an average of 26% of the levels of expression of mRNA encoded by the alternative haplotypes.

When we compared *NrCAM* mRNA levels in the cerebral cortex, midbrain, and hippocampal samples from individuals who were heterozygous for this 5′ *NrCAM* haplotype to samples from individuals who lacked it, expression was about 40% lower in the brains of individuals who displayed the disease-associated haplotype. None of the adjacent genes' expression patterns revealed such evidence for haplotype-specific differential expression.

The core 5′ *NrCAM* haplotype that provided the largest effects on differential expression provided associations that were significant in Sample 1 ($P = 0.0003$), Sample 2 ($P = 0.02$), and Sample 3 ($P = 0.05$).

We then asked if mice with altered levels of NrCAM expression display differences in conditioned place preference, which provides a relatively robust test for alterations in the reward and reward memories induced by abused substances. Heterozygous and/or homozygous NrCAM-knockout mice displayed striking reductions in their preferences for the places where they received either morphine or methamphetamine, compared with wild-type control mice. By contrast, mice of all genotypes displayed similar sensitivities to the acute locomotor stimulant properties of all three drugs.

These data thus provide convergent results from searches for drug-regulated genes and from association-based genome scans for drug abuse vulnerability alleles. The data support the idea that *NrCAM* and *NRXN3* haplotypes contribute to human individual differences in addiction vulnerability in ways that are likely to depend on differences in levels of NrCAM expression and/or regulation and NRXN3 splicing.

A Possible Underlying Neurobiology: Connectivity Constellations

Many brain-based phenotypes and disorders are accompanied by gross brain anatomic

changes or differences compared with brains from individuals without such phenotypes. These anatomical differences provide foci for many pathophysiological inquiries. However, in many other disorders and phenotypes, brains do not reveal reproducible gross pathologies. In these circumstances, investigations more often focus on molecular and biochemical pathways. Pharmacological mechanisms can even provide foci for exploration, when drugs that ameliorate disease symptoms are discovered serendipitously. Roles for mechanisms involved in phenotypes with no gross pathology can be inferred from a number of disorders in which gross brain pathologies have been identified. Cognitive abilities at baseline, accompanied by no gross pathology, can play substantial roles in individual differences in vulnerability to dementing illnesses with gross pathological brain changes, such as AD.[227]

Most of the common brain phenotypes and disorders that lack gross neuropathological underpinnings have been shown to represent complex disorders (or phenotypes) from a genetic perspective.[228] Twin studies support heritability of at least half of the total vulnerability to many such disorders or phenotypes. Major psychiatric disorders such as bipolar disorder, major traits such as general cognitive abilities, and major brain phenotypes such as the volume of frontal and temporal cerebral lobes each display at least 50% heritability in a number of well-performed twin studies (Table 3). Linkage studies, which have been good at identifying individual genes whose variants exert substantial effects on phenotypes or diseases, have established a lack of genes of major effect for most of these phenotypes or disorders. GWA approaches, however, are now revealing more and more of the gene loci that contain variants that contribute to such disorders, and thus more and more of the classes into which genes that contribute to these disorders or traits fall.

Remarkably, virtually all of the brain-based disorders and traits for which GWA data from a variety of laboratories can now be analyzed, as noted above (Table 1), appear to receive substantial contributions from variants in cell adhesion–related genes.

TABLE 3. Connectivity Constellation of Phenotypes and Disorders, each with Sizable Heritable Influences

A. Phenotypes and disorders for which GWA data support involvement of connectivity constellation gene variants in human interindividual differences:
- Addiction (vulnerability to substance dependence)[3,12,33,34,37]
- Alzheimer's disease (reanalyses of data in 4, 5, 300)
- Bipolar disorder (reanalyses of data in 1, 2)
- Brain volume (frontal lobe, Uhl *et al.*, submitted, reanalyses of data in 8, 117, 121, 154)
- Cognitive ability (Uhl *et al.*, submitted)
- Memory[34,36]
- Neuroticism (reanalysis of data in 228)
- Success at abstaining from smoking[301,302]

B. Examples of other phenotypes and disorders that are strong candidates for roles for variants in connectivity constellation genes:
- Antisocial personality disorder
- Attention deficit hyperactivity disorder
- Autism
- Depression
- HIV-associated dementia
- Functional recovery from stroke
- Functional recovery from traumatic brain injury
- Functional recovery from relapses of multiple sclerosis
- Parkinsonian dementia
- Schizophrenia (especially negative symptoms)

Could these repeated observations come simply from unexpectedly large representation of such genes in the genome? We have recently performed bioinformatic searches in which we attempted to determine the number of CAMs. We constructed and performed these searches in ways that were independent of the GWA data described here. These searches used a combination of natural language and keyword-based literature and database searches with DNA sequence motif searches using sequence elements common to CAM families (C.Y. Li, Q.R. Liu, and G.R. Uhl, in preparation). Briefly, we (C.Y. Li) integrated Gene Ontology annotations, domain structure information, and keyword queries of National Center for Biotechnology

Information Entrez Gene annotations in ways that were independent of the hand-annotated lists of cell adhesion–related genes identified by GWA. First, we extracted 196 genes that encoded 281 proteins using the gene ontology term, "cell adhesion (GO:0007155)."[229] Second, we extracted detailed domain features for six subfamilies of CAMs, including cadherins, IgCAMs (CAMS with immunoglobulin domains), integrins, neurexins, catenins, and neuroligins.[230–236] Using these features and Perl scripts, we identified 218 human genes that encoded 532 proteins based on standardized related InterPro domain architectures and proteins mapped onto these architectures.[237] Fourth, we identified lists of cell adhesion genes for the rat and the mouse, as noted above, and searched for human orthologs using homologene.[238] Finally, we identified 1487 results from searches of Entrez Gene using "adhesion AND Homo sapiens [organism]."[238] The second through fourth approaches added 136 additional genes, yielding a total of 496 human CAM genes (C.Y. Li, Q.R. Liu, and G.R. Uhl, in preparation).

With this set of genes in hand, we were now able to estimate the fraction of the genome that they represent. These 496 genes represent 1.6% of the 31,227 human genes now annotated by RefSeq. These 496 genes are often large, occupying a total of 73,695,264 bp of the genome. These sequences thus represent 2.4% of the 3,080,436,051 bp currently elucidated of the human genome sequence. These genes represent 5.8% of the 1,271,259,295 bp currently annotated as gene sequences (e.g., exon, intron, and 10-kb 3′ and 5′ flanking sequence).

We can seek overlap between these CAM genes found through a bioinformatic search and the 99 genes that we have (a) identified by clustered, nominally positive SNPs in one or more of the GWA studies described above and (b) independently (G.R. Uhl) annotated as cell adhesion related. Sixty-four of the 99 genes that we identified as cell adhesion related in the GWA data sets are also included on this independently derived list of "bioinformatic search CAM genes." These 64 overlapping genes represent 12.9% of all of the bioinformatic search CAM genes that we can identify in the genome. These 64 genes total 35,619,485 bp, representing 1.2% of the genome and 2.8% of currently annotated gene sequences.

Despite this relatively modest representation in the genome, then, cell adhesion–related genes are substantially overrepresented in several of the GWA data sets reviewed here when compared with their genomic representation. Cell adhesion–related genes represent 21%–26% of the genes identified by the replicated searches for addiction vulnerability genes in Samples 1 and 2 and Samples 4 and 5. They represent 27% of the genes identified for vulnerability to bipolar disorder, 17% of the genes implicated in individual differences in regional cerebral volumes, 20% of the genes implicated in cognitive abilities, and 20% of the genes implicated in success in smoking cessation.

Further, when we compare the data from the 64 genes identified in both bioinformatic and hand searches with the data for the remaining 35 genes identified through GWA data sets, strong cases can be made for categorizing most of the remaining genes as cell adhesion related. CSMD1 and CSMD2, for example, represent two of the three currently identified cub and sushi multiple-domain genes, although the literature support for cell adhesion functions of these genes is not sufficiently strong to include them on the list of bioinformatic search CAM genes. Reviews and original research papers refer to receptor protein tyrosine phosphatases as CAMs, in ways that correspond to our hand annotation but not to their identification as CAMs in the independent bioinformatic searches.[239–243]

Quality and/or Quantity of Connections Involved in Vulnerabilities

Quantitative morphometric studies have documented the large fraction of volume of cerebral cortical grey matter that is composed of neuropil neuronal processes and their supporting and ensheathing glial elements.[244,245] By comparison, vasculature and

neuronal cell bodies contribute much less to the volume of grey matter. The ways in which neuropil and the neuronal connections that it contains (a) develop, (b) change with experience (or disease), and (c) change with aging all provide interesting avenues for the functional impact of individual differences in neuronal connections.

Heritability estimates are available for several of the human phenotypes that are likely to be influenced by the quality and/or quantity of such neuronal connections. Regional brain volumes have been among the most studied using classical twin methods.[15] Some of the best data currently available come from studies of the volumes of the cerebrum, frontal lobe, temporal lobe, and hippocampus. These twin studies support heritable interindividual differences in these regions that range from 0.4 to 0.6.[15] Because the majority of the volumes of the grey matter components of these regions constitutes connection-rich neuropil, it seems highly likely that differences in the quantity of connections provides one substantial source of these heritable interindividual differences.

Twin methods also provide data that link these heritable determinants of brain volume to heritable determinants of cognitive functions, including memory and estimates of different aspects of cognitive function.[246,247] In one of the largest available twin data sets, virtually all of the heritable individual differences in brain regional volumes could be attributed to genetic influences shared with influences on cognitive abilities (measured primarily by memory tasks).[247] In another twin data set, the overwhelming majority of genetic influences on frontal lobe brain volume were shared with those that influence cognitive ability as measured primarily by tests of executive function.[117]

These data strongly imply roles for individual differences in the quantity of neuronal connections in heritable determinants of cognitive function. No human data of which we are aware directly address qualitative differences in neuronal connections. However, it seems unlikely that genes that influence quantity would fail to influence the quality of connections as well. Interrelations between neurons and glia are also likely to provide significant contributions to the anatomic coherence of white matter pathways. It seems likely that the interindividual differences in diffusion tensor imaging signals that come from white matter differences and that can readily be observed in humans[248] will also receive contributions from "connectivity constellation" genes, to the extent that these individual differences are heritable.[249,250]

Rodent models provide support for quantitative and qualitative effects of variations in connectivity constellation genes. Hints of redundancies in the ways in which these genes might work come from studies of the effects of gene-by-gene interactions in the brains of knockout mice. Knockout mice lacking the addiction-associated gene NrCAM and another immunoglobulin and fibronectin domain CAM gene, L1, display substantial reductions in brain volume and substantial developmental brain-based phenotypes (NrCAM L1KO). However, no observable differences between the brains of wild-type mice and knockouts lacking either NrCAM or L1 alone can be identified. Although gross brain differences in knockouts lacking only a few of the connectivity constellation genes can be identified, it appears likely that more subtle, more qualitative differences might be induced by variants at many more single loci from this list. Larger quantitative differences that can be grossly observed may require contributions from variants at several of these gene loci.

Overlapping Effects of Connectivity Constellation Gene Variants, Comorbidities, and Co-Occurrence Frequencies

Variants in a limited number of genes may contribute to genetic components of susceptibility to a variety of phenotypes. From this hypothesis, one might predict that these phenotypes would occur together more often than expected by chance. The magnitude of co-occurrence, of course, would be expected

to differ based on differential penetrance of each locus on each phenotype as well as other factors.

Relatively strong evidence supports the co-occurrence of vulnerability to a number of the phenotypes listed in Table 3. Bipolar disorder and addiction occur together much more frequently than we would expect by chance alone; almost three-quarters of individuals with bipolar disorder also manifest substance use disorders, according to a recent review and elsewhere.[146,147,251] The co-occurrence of these two heritable phenotypes fits with the substantial overlap that we documented above in comparisons between GWA results from addiction and bipolar disorder. Many of the genes that receive substantial support from both of these studies encode CAMs and other connectivity constellation genes.

As noted above, twin study data support substantial shared overall genetic influences on cognitive abilities and brain volumes.[246,247] Comparison of GWA data sets for these two traits documents substantial overlap that is much greater than would be expected to occur by chance; again, this overlap identifies connectivity constellation genes. Baseline cognitive ability—as reflected by individual differences in educational attainment, for example—is one of the strongest risk factors for AD that can be identified in the literature.[252,253] We have identified strong overlap between GWA results for cognitive ability and those for AD, including a number of connectivity constellation genes.

Neuroticism is a heritable personality trait[254] that differs most in comparisons of addicted versus control individuals.[255,256] We have also identified a strong overlap between GWA data sets for neuroticism and those for addiction[121] (C. Johnson, in preparation). Again, many connectivity constellation genes are identified by both of these approaches.

As increasingly dense GWA data sets for disorders such as antisocial personality disorder and schizophrenia become available, it will be interesting to seek such overlaps with other connectivity constellation disorders.

Contributions of Connectivity Constellation Genes to Morbidity and Mortality from Brain Disorders in the United States

One way to assess the impact of complex polygenic effects on brain disorders is to estimate U.S. costs associated with the disorder, estimate heritabilities from twin study data, and estimate Mendelian and oligogenic contributions from family study, segregation, linkage, and GWA analyses, thus identifying the overall impact of polygenic contributions. We have estimated that the polygenic influences on addiction vulnerability were responsible for $212.20 billion in costs to the United States in 2004.[257–259] Other connectivity constellation disorders and the relevant estimates of costs attributable to polygenic contributions include (a) AD and dementias (with perhaps $74 billion in such costs); (b) pain and migraine ($58.8 billion); (c) anxiety disorders ($24 billion), schizophrenia ($39 billion); (d) depressive illnesses (almost $19 billion); (e) developmental disorders (with perhaps $5 billion in such costs); (f) stroke ($2 billion); (g) Parkinson's disease ($1.4 billion); (h) multiple sclerosis ($3 billion); and (i) seizures ($0.6 billion). If variants in the connectivity constellation of genes contribute 20% to the complex genetic impact of these disorders, then the costs attributable to cell adhesion–related gene variants could amount to almost $90 billion in the United States annually.

Classes of Genes Identified in Multiple GWA Samples for Multiple Phenotypes: Focus on "Druggable" Targets

The emphasis that we have placed on CAM genes here should not obscure the identification of a large number of genes that encode proteins that are more classically "druggable." In particular, enzymes, receptors and ligands (G protein–coupled receptors, ligand-gated ion

channels, and peptide ligands), ion channels, and transporters provide targets for many currently effective drugs and are thus thought of as more druggable than CAMs, for which there is less precedent for efficacious small-molecule therapeutics. It is therefore instructive to identify the potential targets derived from genes in these families that are supported by results in replicated GWA studies of at least three of the phenotypes assessed in this review. It is worth noting that many of these drug targets in fact can be targeted by promising lead compounds that have already been identified (see below).

"Druggable" Genes That Are Most Strongly Supported by the Current Analyses

Cumulative nominal *P* values, based on the strongest clusters on nominally positive SNPs in each of these individual genes in studies of a variety of different phenotypes, provide support for individual genes that is as striking as 10^{-3} to less than 10^{-5} for several of these druggable genes.

Enzyme-related genes are the most numerous group, with PRKG1, CAMK1D, CHN2, FHIT, and SERPINA1 providing the most significant convergent observations. PRKG1, the cyclic guanosine 5′-monophosphate (cGMP)-dependent protein kinase 1, and FHIT, the fragile histone triad gene that represents a diadenosinase, each displayed clustered, nominally positive SNPs in several of these studies in ways that make them especially unlikely to represent chance findings. PRKG1 is expressed in the brain in hippocampal, cerebellar, and other neurons.[260] Nitric oxide dramatically modulates brain cGMP systems; PRKG1 thus provides a major target for the products of nitric oxide synthases (NOS). Mnemonic and addictive functions can each be altered by changes in cGMP-dependent protein kinase and/or NOS.[261] Identification of potent imidazopyridine PRKG inhibitors for use in lower species could well provide leads for the development of similar compounds for use in humans.[262]

FHIT was initially named the fragile histone triad gene based on its location at the chromosome 3p Fra3B locus, which provides one of the most common neoplasia-associated human genomic fragile sites. FHIT is highly expressed in the brain, with robust expression localized in regions that include the hippocampus. The high levels of brain expression of FHIT protein and its characterization as the major human enzyme that can hydrolyze diadenosine-polyphosphates provide a strong link to purinergic signaling.[263] Diadenosine polyphosphates are released from synaptic vesicles along with more classical transmitters. They can activate P2Y(1), P2Y(2), and P2Y(4) receptors; homomeric P2X(1), P2X(2), P2X(3), P2X(4), and P2X(6) receptors; and even a P4 receptor-operated Ca^{++} channel.[261] Recent reports have also implicated FHIT in a nonenzymatic function, altering the expression of the cell adhesion–related protein beta catenin.[264] Robust inhibition of FHIT's diadenosine-triphosphatase activity by suramin provides a starting place for developing more selective inhibitors.[265,266] However, recent identification of FHIT as a modulator of beta catenin–induced transcriptional regulation may also place this gene in the cell adhesion signaling cascade.[263]

CAMK1D is expressed in neurons, including those of the hippocampus, at relatively high levels. Although little direct information about the functional effects of its brain expression is available, other calcium/calmodulin-dependent protein kinases play large roles in memory-like processes. Useful reagents to study CAMK1D activities in peripheral systems might well come from studies of mimics of CAMK1D sequences.[267–269]

LARGE encodes a likely enzyme; this gene also receives much cumulative support from these GWA data sets. LARGE encodes a protein with high homology to acetylglucosaminyltransferases. The brain provides many of the LARGE-expressed sequence tags, and *in situ* hybridization has documented high levels of expression in apparent neurons in a number

of brain regions, including the hippocampus and cortex.[172] In muscle tissue, interrelationships between dystroglycans and LARGE and its family members have been so well documented that interactions with other dystroglycans in the brain also appear likely.[270] Humans with LARGE mutations can express developmental delay as well as muscular dystrophies; a spontaneously occurring mouse model is also available.[271]

The phosphodiesterase PDE4D, which also receives substantial support, is an enzyme expressed in neurons and other cells in the brain in a number of regions that include the hippocampus. PDE4D knockouts display altered results in antidepressant testing.[272] Variants at the PDE4D locus have been associated with a number of additional human phenotypes. For example, PDE4D SNPs have been associated with neuroticism in the initial analyses of data from Sample 19.[9] A number of studies, but not all, have linked and/or associated PDE4D haplotypes with an altered risk of stroke.[273] A PDE4D SNP has been associated with sleepiness in Framingham study subjects in a 100,000-SNP GWA study.[274] Rolipam and other PDE4D inhibitors or stimulators have been extensively characterized, providing starting points for synthesis of more selective drugs.[275] Memory-enhancing effects of PDE4 inhibitors have been documented in rodent behavioral assays and provide relatively straightforward links with the GWA data that we review here.[12]

RYR3 is the channel gene that is most heavily supported by these data. RYR3 is a site at which calcium-activated calcium efflux from sequestered intracellular stores elevates free cytoplasmic calcium levels. Brain RYR3 expression is highest in the hippocampus.[276,277] Significant evidence links RYRs to associative learning.[278] RYR3 knockout mice display alterations in several sorts of learning tasks.[279] Ryanodine and a number of other small molecules appear to provide excellent starting points for synthesis of selective RYR3 inhibitors.[280]

Other genes on this list of classically druggable targets that display substantial cumulative nominal *P* values include genes that encode a peptide or growth factor ligand, G protein–coupled receptors, and transporters.

FGF14 isoforms are widely expressed in a number of brain regions during development and are expressed in a more focal fashion in the adult brain.[281] FGF14 lacks signal sequences and may thus play intracellular roles that include interactions with ion channels and intracellular signal-transducing proteins.[282] The FGF14 gene is a candidate to play roles in the control of midbrain dopamine neurons, based on mouse strain comparison studies.[283,284] FGF14 expression is upregulated by methylphenidate.[285] FGF14 knockout mice display substantial alterations in learning and aspects of long-term potentiation.[286] Humans with FGF14 mutations also display cognitive changes.[121] Although we are aware of no small-molecule FGF14 agonists or antagonists, such development might well benefit from the substantial progress in producing small-molecule peptidomimetics in other fields.

Recent progress in identifying ligands for G protein–coupled glutamate receptors (see Caroll *et al.*, this volume) provides substantial hope that mGluR7 ligands will soon be available to test roles for this receptor in substance dependence and other brain functions. The high significance attained by ABCC4 also supports strong consideration of drugs targeting this interesting ABC cassette transporter.

The genes discussed here provide examples of potential therapeutic targets; many of the other genes identified by these studies may also provide equally tractable drug target sites. Indeed, some of the genes that are identified in genome-wide data from only one or two phenotypes might conceivably provide targets for drugs with more favorable specificities and therapeutic indices than genes identified by GWA data for many phenotypes.

Directions, Summary, and Conclusions

It is an exciting time to be able to summarize and review the rapidly emerging data on the complex genetics of human addiction vulnerability and related phenotypes. GWA results for dependence on several different classes of addictive substances converge with each other in striking fashion that is highly unlikely to be attributable to chance. Studies of dependence phenotypes in samples of individuals from several different racial and ethnic backgrounds support the idea that many of the allelic variants that predispose individuals to these common disorders are so evolutionarily old that they are present in members of each major current human population. These data, combined with the varying results from linkage-based studies, fit a genetic architecture for addiction that is based on polygenic contributions from common allelic variants. Such a genetic architecture is quite consistent with data from family, adoption, and twin classical genetic studies.

The identification of genes with markers whose allelic frequencies distinguish addicts of several different ethnicities from matched controls supports a "common disease/common allele" genetic architecture[287] for at least much of addiction vulnerability. The convergent data derived from studies of individuals with addictions to substances in several different pharmacological classes supports the idea that higher-order pharmacogenomic and pharmacogenetic variations enhance vulnerability to many addictions. These results do not exclude additional contributions to addiction vulnerability from genomic variants that influence vulnerability to specific substances or variants that are found only in specific populations. Nevertheless, the findings presented here provide promise for enhancing understanding of features that are common to human addictions in ways that could facilitate efforts to personalize prevention and treatment strategies for debilitating addictive disorders.

Identification of addiction-associated variants in genes that are likely to alter the quality of brain connections provides a first step toward defining a new neurobiology for the underpinnings of specific diseases and phenotypes. For many of these diseases and phenotypes, very little current research focuses on the direct study of brain connections. The *connectivity constellation* concepts that we introduce here support studies that develop and use current and novel means for assessing the qualities and quantities of brain connections, especially in contexts in which they assess their functional properties. We have identified contributions of connectivity constellation genes to volumes of the same brain regions in which many of these genes are expressed. This convergence may provide new insights into data that document individual differences in frontal lobe volume and/or in function, detected by volumetric, deoxyglucose PET and/or fMRI imaging, for virtually all of the connectivity constellation phenotypes or disorders noted here.[8,117]

The addiction vulnerability genes identified in this work contribute to the growing body of data that implicate cell adhesion and related memory-like and other cognitive processes in addiction. Studies that alter reconsolidation and other memory-related processes using knockout mice, protein synthesis inhibitors, and/or pharmacologic treatments demonstrate powerful influences on addictions.[288,289] This empirical evidence enriches theoretical work that increasingly recognizes memory-like features for addiction[36] and work that implicates memory-associated brain regions in relapse to addiction. Such work also complements clinical observations that document that addicts' enhanced vulnerabilities to substance abuse relapse can persist for decades after their last prior use of addictive substances.

Substantial evidence also exists for the generalization of these results from addiction. This evidence comes from the significant overlaps between the molecular genetics of addiction and the molecular genetics of a number of

related phenotypes and disorders. Overlap with bipolar disorder provides one of several likely psychiatric diagnoses for which shared genetic influences are likely *a priori*, based on the substantial heritabilities of both addiction and bipolar disorder as well as the high frequency of addiction–bipolar disorder comorbidity.[146,147] This same logic suggests that abundant shared genetics may well also underpin the frequent comorbidities between addictions and antisocial personality and conduct disorders.[251] Less compelling evidence points to overlaps with other depressive, anxiety, and schizophrenic disorders as well.[251]

We have sought evidence for genetic influences that are shared between addiction and (a) frontal lobe brain volumes, (b) cognitive function, and (c) AD. Hypotheses about such shared genetic influences are based, in part, on initial observations that so many of the genes that we and others have identified in addiction GWA relate to cell connections. These molecularly based hypotheses were reinforced by the evidence for substantial, complex genetic components of each of these phenotypes. These hypotheses were also strengthened by evidence, although often from small samples, that appears to document (a) small frontal lobe volumes in samples of addicts,[19,290] (b) lower performance levels on tests of cognitive and executive function in samples of addicts,[291–293] and (c) a large role for heritability versus a small role for the drug exposure itself in determining the cognitive abilities of twin pair members who are discordant for cannabis use.[294] These hypotheses are further reinforced by (a) twin data that document strong shared genetic influences on frontal brain volumes and cognitive function measures,[246,247] (b) smaller head (and thus likely brain) sizes in individuals who go on to develop AD decades later,[295] and (c) lower levels of educational attainment (and thus cognitive function) in individuals who go on to develop AD many years later.[252,253] Little convincing prior evidence linked addiction liability directly with AD, although reasonably strong links have been made between addiction vulnerability and the related, often-dementing neurodegenerative illness Parkinson's disease[296] (further reviewed in 297). Identification of the genes documented here should lead to the elucidation of features of addiction vulnerability that might arise preferentially from the connectivity constellation gene variants that also alter vulnerability to AD.

The focus on genes in the current analyses should not obscure the fact that many of the SNPs and loci identified in each of these GWA studies lie between annotated genes. As the biological roles for intergenic regions become better understood generally, it should be increasing.

Disease-associated markers, both within and between genes, can begin to allow us to assess individual differences in vulnerability to addiction based on profiles of genotypes. In settings in which prevention of addiction is sought, addiction vulnerability genomic profiles could help to target more (or different) prevention resources to individuals at the greatest (or at different) genetic risk. When a therapeutic opiate is considered for chronic, noncancer pain, for example, the costs of engendering substance dependence are likely to be sufficient to justify genotyping, even if the results provide only partial information about risk assessment and minimization for prescribing physicians. When treatment for an established dependence on nicotine, opiates, or alcohol is contemplated, a number of different therapeutic options with different pharmacological mechanisms of action are now available.[298] Subsets of the SNPs that we have associated with success in quitting smoking appear to selectively influence success in responding to bupropion, whereas others appear to selectively influence success in response to nicotine replacement. Replication and extension of these observations to treatments for alcohol, opiates, and other addictive substances will increase the likelihood that SNP markers will aid the personalization of antiaddiction therapies in the near future in ways that are now affecting the design of clinical trials in this area.

The shared genetic influences on addiction and these other brain-based phenotypes, as well as our implication of the roles for variants in cell adhesion genes in many of these phenotypes, underscore the likelihood that the biology of these systems will demand improved approaches to understanding more of the details of the estimated approximately 10^{14} synaptic connections in the cerebral cortex and those in other areas.[299] Gross phenotypes, such as lobar brain volumes, are likely to vastly underrepresent the true complexity of the qualitative differences that are driven by these gene variants and provide only weak reporters for quantitative differences. The present genetic analyses underscore the urgency of developing better approaches to elucidating individual differences in connectivities in the human brain, both living and postmortem.

As we elaborate and extend these molecular genetic approaches, it seems likely that the number of genes with variants that contribute to individual differences in human vulnerabilities to addiction and related phenotypes will continue to grow, and that many of the genes listed here will receive additional support. By the time that this review is printed, more detailed data from arrays that contain markers for almost 1 million SNP and 1 million copy number variant probes will enhance the picture that we present here (Liu *et al.*, in preparation). These studies will add power to detect differences in many of the genes and intragenic regions that are not well studied using the 500–600K data sets analyzed in this review, and to allow us to understand the potential roles of copy number variation in these addiction and related phenotypes. More and more such studies will provide an increasing ability to better understand, better prevent, and better treat these common and debilitating conditions.

This work, taken together, supports the idea that the heritable brain bases for individual differences in addiction vulnerability lie squarely in the midst of the repertoire of common complex determinants of individual differences that are manifested in many heritable, complex brain disorders and phenotypes. Such conclusions place the biology of addictions squarely in the midst of important biologies of a number of brain phenotypes and disorders, in ways that we hope will benefit them all.

Acknowledgments

We thank the subjects for each of these samples. We thank collaborators who include H. Ujike and the JGIDA methamphetamine investigators, S.K. Li and the Taiwan methamphetamine investigators, Jed Rose, Caryn Lerman, Ray Niaura, Sean David, Gary Swan, and Christina Lessov-Schlaggar. We are grateful to TGen, GlaxoSmithKline (GSK), Framingham study investigators, the NicSNP group, and the Wellcome Trust Case Control Consortium for access to genotype data analyzed here. Rachel Gibson, S. Seshadri, and J. Pollock were of particular help in obtaining access to the GSK, Framingham, and NicSNP data sets cited here. We are grateful for dedicated help with clinical characterization of NIDA subjects from Dan Lipstein, Fely Carillo, Carlo Contoreggi, Fred Snyder, and other Johns Hopkins–Bayview support staff. We benefited from passionate discussions of statistical issues with Dr. Daniel Naiman and from the Baltimore Epidemiologic Catchment Area follow-up study that was generously provided by Dr. J. Anthony. We thank NHLBI staff and the twin study technicians and investigators, P.A. Wolf, B.L. Miller, T. Reed, L. Epstein, L. Hawk, P. Shields, F. Patterson, A. Pinto, M. Rukstalis, W. Berrettini, R. Brown, E. Richardson, F. M. Behm, P. Kukovich, E.C. Westman, and G. Samsa for their rigor in overseeing data collection at their research sites. We acknowledge financial support from NIH–IRP (NIDA), U.S. Department of Health and Human Services, and are also grateful for support for some of the studies discussed in detail here from the Taiwan and Japanese Ministries for Science and Technology, NIH grants P50CA/DA84718, RO1CA 63562, HL32318, DA08511, P50CA84719, 1K08 DA14276–05, and HL51429, support from the Welcome

Trust (076113), the Pennsylvania Department of Health (which specifically disclaims responsibility for any analyses, interpretations, or conclusions), GlaxoSmithKline, Inc., and unrestricted support for studies of adult smoking cessation from Phillip Morris USA. Some human brain tissues were obtained from the Brain and Tissue Bank for Developmental Disorders supported through NIH contract NO1-HD-1-3138.

Conflicts of Interest

The authors declare no conflicts of interest.

Glossary

A priori: Existing in the mind prior to and independent of experiments.

Balancing selection: A natural process that results in the survival and reproductive success of individuals or groups best adjusted to their environment and that leads to the perpetuation of genetic qualities best suited to that particular environment.

Between-locus heterogeneity: A single disorder, trait, or pattern of traits caused by mutations in genes at different chromosomal loci.

Common disease and common allele model: A model proposing that an illness results from the cumulative impact of multiple common, small-effect, genetic variants, interacting with environmental exposures to exceed a biological threshold.

Complex genetic phenotype (polygenic and multifactorial traits): Any *phenotype* that results from the effect of multiple *genes* at two or more loci, with possible environmental influences too.

Epigenetic: Changes in the regulation of the expression of gene activity without alteration of the DNA sequence.

Epistasis: A process in which a mutation in one gene alters the expression of a different gene.

Genetic heterogeneity: A single disorder, trait, or pattern of traits caused by genetic factors in some cases and nongenetic factors in others.

Genetic selection: Differential and nonrandom reproduction of different genotypes, operating to alter the gene frequencies within a population.

Genome-wide association study: Any study of genetic variation across the entire human genome that is designed to identify genetic associations with observable traits (such as blood pressure or weight) or the presence or absence of a disease or condition.

Linkage: The tendency for genes or segments of DNA closely positioned along a chromosome to segregate together at meiosis and therefore be inherited together.

Linkage analysis: Testing DNA sequence polymorphisms that are near or within a gene of interest to track within a family the inheritance of a disease-causing variant in a given gene.

Linkage disequilibrium: In a population, the co-occurrence of a specific DNA marker and a disease at a higher frequency than would be predicted by chance.

Pharmacogenomics: The study of how an individual's genomic system affects the body's response to drugs.

Pharmacogenetics: The study of specific genes, such as drug-metabolizing enzymes.

Pleiotropy: Multiple, often seemingly unrelated, physical effects caused by a single altered gene or pair of altered genes.

Segregation analysis: The determination of the number of progeny that have inherited distinct and mutually exclusive phenotypes.

Susceptibility gene: A gene mutation that increases the likelihood that an individual will develop a certain disease or disorder. When such a mutation is inherited, the development of symptoms is more likely but not certain.

Transitive: Passing over to or affecting something else.

Within-locus heterogeneity: A single disorder, trait, or pattern of traits influenced by several different variants at a single chromosomal locus.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table SI. Additional discussion of the derivation of Gene Detective
Table SII. Additional table listing "non-detectable" genes

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

References

1. Wellcome Case Control TrustConsortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447:** 661–678.
2. Baum, A.E. *et al.* 2008. A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder. *Mol. Psychiatry* **13:** 197–207.
3. Bierut, L.J. *et al.* 2007. Novel genes identified in a high-density genome wide association study for nicotine dependence. *Hum. Mol. Genet.* **16:** 24–35.
4. Coon, K.D. *et al.* 2007. A high-density whole-genome association study reveals that APOE is the major susceptibility gene for sporadic late-onset Alzheimer's disease. *J. Clin. Psychiatry* **68:** 613–618.
5. Li, H. *et al.* 2008. Candidate single-nucleotide polymorphisms from a genomewide association study of Alzheimer disease. *Arch. Neurol.* **65:** 45–53.
6. Dudbridge, F. *et al.* 2006. Detecting multiple associations in genome-wide studies. *Hum. Genomics* **2:** 310–317.
7. Frayling, T.M. 2007. Genome-wide association studies provide new insights into type 2 diabetes aetiology. *Nat. Rev. Genet.* **8:** 657–662.
8. Seshadri, S. *et al.* 2007. Genetic correlates of brain aging on MRI and cognitive test measures: A genome-wide association and linkage analysis in the Framingham Study. *BMC Med. Genet.* **8**(Suppl 1): S15.
9. Gottlieb, D.J. *et al.* 2007. Genome-wide association of sleep and circadian phenotypes. *BMC Med. Genet.* **8**(Suppl 1): S9.
10. Uhl, G.R. *et al.* 2001. Polysubstance abuse-vulnerability genes: Genome scans for association, using 1,004 subjects and 1,494 single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **69:** 1290–300.
11. Liu, Q.R. *et al.* 2005. Pooled association genome scanning: Validation and use to identify addiction vulnerability loci in two samples. *Proc. Natl. Acad. Sci. USA* **102:** 11864–11869.
12. Liu, Q.R. *et al.* 2006. Addiction molecular genetics: 639,401 SNP whole genome association identifies many "cell adhesion" genes. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **141:** 918–925.
13. Hung, R.J. *et al.* 2008. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature* **452:** 633–637.
14. Thorgeirsson, T.E. *et al.* 2008. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* **452:** 638–642.
15. Peper, J.S. *et al.* 2007. Genetic influences on human brain structure: A review of brain imaging studies in twins. *Hum. Brain Mapp.* **28:** 464–473.
16. Sim, M.E. *et al.* 2007. Cerebellar gray matter volume correlates with duration of cocaine use in cocaine-dependent subjects. *Neuropsychopharmacology* **32:** 2229–2237.
17. Brody, A.L. *et al.* 2004. Differences between smokers and nonsmokers in regional gray matter volumes and densities. *Biol. Psychiatry* **55:** 77–84.
18. Thompson, P.M. *et al.* 2004. Structural abnormalities in the brains of human subjects who use methamphetamine. *J. Neurosci.* **24:** 6028–6036.
19. Franklin, T.R. *et al.* 2002. Decreased gray matter concentration in the insular, orbitofrontal, cingulate, and temporal cortices of cocaine patients. *Biol. Psychiatry* **51:** 134–142.
20. Baicy, K. & E.D. London. 2007. Corticolimbic dysregulation and chronic methamphetamine abuse. *Addiction* **102**(Suppl 1): 5–15.
21. Tomasi, D. *et al.* 2007. Thalamo-cortical dysfunction in cocaine abusers: Implications in attention and perception. *Psychiatry Res.* **155:** 189–201.
22. Guerrini, I. *et al.* 2005. Genetic linkage analysis supports the presence of two susceptibility loci for alcoholism and heavy drinking on chromosome 1p22.1–11.2 and 1q21.3–24.2. *BMC Genet.* **6:** 11.
23. Zhong, X. & H. Zhang. 2005. Linkage analysis and association analysis in the presence of linkage using age at onset of COGA alcoholism data. *BMC Genet.* **6**(Suppl 1): S31.
24. Clarimon, J. *et al.* 2007. Linkage disequilibrium and association analysis of alpha-synuclein and alcohol

and drug dependence in two American Indian populations. *Alcohol Clin. Exp. Res.* **31:** 546–554.
25. Gelernter, J. *et al.* 2007. Genomewide linkage scan for nicotine dependence: Identification of a chromosome 5 risk locus. *Biol. Psychiatry* **61:** 119–126.
26. Hopfer, C.J. *et al.* 2007. A genome-wide scan for loci influencing adolescent cannabis dependence symptoms: Evidence for linkage on chromosomes 3 and 9. *Drug Alcohol Depend.* **89:** 34–41.
27. Kuo, P.H. *et al.* 2006. Identification of susceptibility loci for alcohol-related traits in the Irish Affected Sib Pair Study of Alcohol Dependence. *Alcohol Clin. Exp. Res.* **30:** 1807–1816.
28. Li, M.D. *et al.* 2007. Linkage and association studies in African- and Caucasian-American populations demonstrate that SHC3 is a novel susceptibility locus for nicotine dependence. *Mol. Psychiatry* **12:** 462–473.
29. Loh, E.W. *et al.* 2007. Association analysis of GABA receptor subunit genes on 5q33 with heroin dependence in a Chinese male population. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **144:** 439–443.
30. Pomerleau, O.F. *et al.* 2007. Genome-wide linkage analysis for smoking-related regions, with replication in two ethnically diverse populations. *Nicotine Tob. Res.* **9:** 955–958.
31. Saccone, S.F. *et al.* 2007. Genetic linkage to chromosome 22q12 for a heavy-smoking quantitative trait in two independent samples. *Am. J. Hum. Genet.* **80:** 856–866.
32. Uhl, G.R. *et al.* 2008. "Higher order" addiction molecular genetics: Convergent data from genome-wide association in humans and mice. *Biochem. Pharmacol.* **75:** 98–111.
33. Johnson, C. *et al.* 2006. Pooled association genome scanning for alcohol dependence using 104,268 SNPs: Validation and use to identify alcoholism vulnerability loci in unrelated individuals from the collaborative study on the genetics of alcoholism. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **141:** 844–853.
34. Uhl, G.R. *et al.* 2007. Molecular genetics of nicotine dependence and abstinence: Whole genome association using 520,000 SNPs. *BMC Genet.* **8:** 10.
35. Saccone, S.F. *et al.* 2007. Cholinergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3713 SNPs. *Hum. Mol. Genet.* **16:** 36–49.
36. Uhl, G.R. *et al.* 2008. Molecular genetics of successful smoking cessation: Convergent genome-wide association results. *Arch. Gen. Psychiatry* **65:** 683–690.
37. Uhl, G.R. *et al.* 2008. Genome-wide association for methamphetamine dependence: Convergent results from two samples. *Arch. Gen. Psychiatry* (in press).
38. Meaburn, E. *et al.* 2005. Genotyping DNA pools on microarrays: Tackling the QTL problem of large samples and large numbers of SNPs. *BMC Genomics* **6:** 52.
39. Meaburn, E. *et al.* 2006. Genotyping pooled DNA using 100K SNP microarrays: A step towards genomewide association scans. *Nucleic Acids Res.* **34:** e27.
40. Butcher, L.M. *et al.* 2005. Association analysis of mild mental impairment using DNA pooling to screen 432 brain-expressed single-nucleotide polymorphisms. *Mol. Psychiatry* **10:** 384–392.
41. Butcher, L.M. *et al.* 2005. SNPs, microarrays and pooled DNA: Identification of four loci associated with mild mental impairment in a sample of 6000 children. *Hum. Mol. Genet.* **14:** 1315–1325.
42. Butcher, L.M. *et al.* 2004. Genotyping pooled DNA on microarrays: A systematic genome screen of thousands of SNPs in large samples to detect QTLs for complex traits. *Behav. Genet.* **34:** 549–555.
43. Sham, P. *et al.* 2002. DNA pooling: A tool for large-scale association studies. *Nat. Rev. Genet.* **3:** 862–871.
44. Pearson, J.V. *et al.* 2007. Identification of the genetic basis for complex disorders by use of pooling-based genomewide single-nucleotide-polymorphism association studies. *Am. J. Hum. Genet.* **80:** 126–139.
45. Uhl, G.R. *et al.* 1995. Genetic influences in drug abuse. In *Psychopharmacology: The Fourth Generation of Progress*. F.E. Gloom & D.J. Kupfer, Eds.: 1793–2783. Raven Press. New York, NY.
46. Tsuang, M.T. *et al.* 1998. Co-occurrence of abuse of different drugs in men: The role of drug-specific and shared vulnerabilities. *Arch. Gen. Psychiatry* **55:** 967–972.
47. Karkowski, L.M. *et al.* 2000. Multivariate assessment of factors influencing illicit substance use in twins from female-female pairs. *Am. J. Med. Genet.* **96:** 665–670.
48. True, W.R. *et al.* 1999. Interrelationship of genetic and environmental influences on conduct disorder and alcohol and marijuana dependence symptoms. *Am. J. Med. Genet.* **88:** 391–397.
49. Merikangas, K.R. *et al.* 1998. Familial transmission of substance use disorders. *Arch. Gen. Psychiatry* **55:** 973–979.
50. Woodward, C.E. *et al.* 1996. Tobacco, alcohol and drug use in 8–16 year old twins. *NIDA Res. Monograph.* **162:** 309.
51. Tsuang, M.T. *et al.* 1996. Genetic influences on DSM-III-R drug abuse and dependence: A study of 3,372 twin pairs. *Am. J. Med. Genet.* **67:** 473–477.
52. Kendler, K.S. & C.A. Prescott. 1998. Cocaine use, abuse and dependence in a population-based sample of female twins. *Br. J. Psychiatry* **173:** 345–350.

53. Kendler, K.S. *et al.* 2006. Illicit psychoactive substance use, abuse and dependence in a population-based sample of Norwegian twins. *Psychol. Med.* **36:** 955–962.
54. Agrawal, A. *et al.* 2004. A twin study of early cannabis use and subsequent use and abuse/dependence of other illicit drugs. *Psychol. Med.* **34:** 1227–1237.
55. Grove, W.M. *et al.* 1990. Heritability of substance abuse and antisocial behavior: A study of monozygotic twins reared apart. *Biol. Psychiatry* **27:** 1293–1304.
56. Gynther, L.M. *et al.* 1995. A twin study of non-alcohol substance abuse. *Psychiatry Res.* **56:** 213–220.
57. Chen, C.C. *et al.* 1999. Interaction between the functional polymorphisms of the alcohol-metabolism genes in protection against alcoholism. *Am. J. Hum. Genet.* **65:** 795–807.
58. Thomasson, H.R. *et al.* 1991. Alcohol and aldehyde dehydrogenase genotypes and alcoholism in Chinese men. *Am. J. Hum. Genet.* **48:** 677–681.
59. Chen, W.J. *et al.* 1996. Alcohol-metabolising genes and alcoholism among Taiwanese Han men: Independent effect of ADH2, ADH3 and ALDH2. *Br. J. Psychiatry* **168:** 762–767.
60. Shen, Y.C. *et al.* 1997. Polymorphism of ADH and ALDH genes among four ethnic groups in China and effects upon the risk for alcoholism. *Alcohol Clin. Exp. Res.* **21:** 1272–1277.
61. Higuchi, S. 1994. Polymorphisms of ethanol metabolizing enzyme genes and alcoholism. *Alcohol Alcohol Suppl.* **2:** 29–34.
62. Higuchi, S. *et al.* 1994. Aldehyde dehydrogenase genotypes in Japanese alcoholics. *Lancet* **343:** 741–742.
63. Higuchi, S. *et al.* 1995. Alcohol and aldehyde dehydrogenase polymorphisms and the risk for alcoholism. *Am. J. Psychiatry* **152:** 1219–1221.
64. Maezawa, Y. *et al.* 1995. Alcohol-metabolizing enzyme polymorphisms and alcoholism in Japan. *Alcohol Clin. Exp. Res.* **19:** 951–954.
65. Nakamura, K. *et al.* 1996. Characteristics of Japanese alcoholics with the atypical aldehyde dehydrogenase 2*2. I. A comparison of the genotypes of ALDH2, ADH2, ADH3, and cytochrome P-4502E1 between alcoholics and nonalcoholics. *Alcohol Clin. Exp. Res.* **20:** 52–55.
66. Tanaka, F. *et al.* 1997. Polymorphism of alcohol-metabolizing genes affects drinking behavior and alcoholic liver disease in Japanese men. *Alcohol Clin. Exp. Res.* **21:** 596–601.
67. Luczak, S.E. *et al.* 2002. Genetic risk for alcoholism relates to level of response to alcohol in Asian-American men and women. *J. Stud. Alcohol.* **63:** 74–82.
68. Schuckit, M.A. & J. Duby. 1982. Alcohol-related flushing and the risk for alcoholism in sons of alcoholics. *J. Clin. Psychiatry* **43:** 415–418.
69. Berrettini, W. *et al.* 2008. alpha-5/alpha-3 nicotinic receptor subunit alleles increase risk for heavy smoking. *Mol. Psychiatry* **13:** 368–373.
70. Porjesz, B. *et al.* 1998. Amplitude of visual P3 event-related potential as a phenotypic marker for a predisposition to alcoholism: Preliminary results from the COGA Project. Collaborative Study on the Genetics of Alcoholism. *Alcohol Clin. Exp. Res.* **22:** 1317–1323.
71. Dick, D.M. *et al.* 2004. Association of GABRG3 with alcohol dependence. *Alcohol Clin. Exp. Res.* **28:** 4–9.
72. Gelernter, J. *et al.* 2004. Results of a genomewide linkage scan: Support for chromosomes 9 and 11 loci increasing risk for cigarette smoking. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **128:** 94–101.
73. Saccone, N.L. *et al.* 2000. A genome screen of maximum number of drinks as an alcoholism phenotype. *Am. J. Med. Genet.* **96:** 632–637.
74. Schuckit, M.A. *et al.* 2001. A genome-wide search for genes that relate to a low level of response to alcohol. *Alcohol Clin. Exp. Res.* **25:** 323–329.
75. Bergen, A.W. *et al.* 2003. Genomic regions linked to alcohol consumption in the Framingham Heart Study. *BMC Genet.* **4**(Suppl 1): S101.
76. Bierut, L.J. *et al.* 2004. A genomic scan for habitual smoking in families of alcoholics: Common and specific genetic factors in substance dependence. *Am. J. Med. Genet. A* **124:** 19–27.
77. Dick, D.M. *et al.* 2006. Endophenotypes successfully lead to gene identification: Results from the collaborative study on the genetics of alcoholism. *Behav. Genet.* **36:** 112–126.
78. Dick, D.M. *et al.* 2002. Suggestive linkage on chromosome 1 for a quantitative alcohol-related phenotype. *Alcohol Clin. Exp. Res.* **26:** 1453–1460.
79. Dick, D.M. *et al.* 2006. Association between GABRA1 and drinking behaviors in the collaborative study on the genetics of alcoholism sample. *Alcohol Clin. Exp. Res.* **30:** 1101–1110.
80. Edenberg, H.J. & T. Foroud. 2006. The genetics of alcoholism: Identifying specific genes through family studies. *Addict. Biol.* **11:** 386–396.
81. Pinnaduwage, D. & L. Briollais. 2005. Comparison of genotype- and haplotype-based approaches for fine-mapping of alcohol dependence using COGA data. *BMC Genet.* **6**(Suppl 1): S65.
82. Porjesz, B. *et al.* 2002. Linkage and linkage disequilibrium mapping of ERP and EEG phenotypes. *Biol. Psychol.* **61:** 229–248.
83. Reck, B.H. *et al.* 2005. Analysis of alcohol dependence phenotype in the COGA families using

covariates to detect linkage. *BMC Genet.* **6**(Suppl 1): S143.
84. Yang, H.C. *et al.* 2005. A genome-wide scanning and fine mapping study of COGA data. *BMC Genet.* **6**(Suppl 1): S30.
85. Gudmundsson, J. *et al.* 2007. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat. Genet.* **39:** 631–637.
86. Plomin, R. *et al.* 1977. Genotype-environment interaction and correlation in the analysis of human behavior. *Psychol. Bull.* **84:** 309–322.
87. Scarr, S. & K. McCartney. 1983. How people make their own environments: A theory of genotype greater than environment effects. *Child Dev.* **54:** 424–435.
88. Posthuma, D. *et al.* 2003. Theory and practice in quantitative genetics. *Twin Res.* **6:** 361–376.
89. Lytton, H. *et al.* 1977. Environmental and genetical causes of variation in ethological aspects of behavior in two-year-old boys. *Soc. Biol.* **24:** 200–211.
90. Falconer, D.S. & T.F.C. MacKay. 1996. *Introduction to Quantitative Genetics*. 4th ed. Longan Group Ltd. London.
91. Buster, M.A. & J.L. Rodgers. 2000. Genetic and environmental influences on alcohol use: DF analysis of NLSY kinship data. *J. Biosoc. Sci.* **32:** 177–189.
92. Stanke, F. *et al.* 2008. Diversity of the basic defect of homozygous CFTR mutation genotypes in humans. *J. Med. Genet.* **45:** 47–54.
93. Douglas, M.R. *et al.* 2007. Genetics of Parkinson's disease and parkinsonism. *Expert Rev. Neurother.***7:** 657–666.
94. Stephan, D.A. 2008. Unraveling autism. *Am. J. Hum. Genet.* **82:** 7–9.
95. Alarcon, M. *et al.* 2008. Linkage, association, and gene-expression analyses identify CNTNAP2 as an autism-susceptibility gene. *Am. J. Hum. Genet.* **82:** 150–159.
96. Arking, D.E. *et al.* 2008. A common genetic variant in the neurexin superfamily member CNTNAP2 increases familial risk of autism. *Am. J. Hum. Genet.* **82:** 160–164.
97. Bakkaloglu, B. *et al.* 2008. Molecular cytogenetic analysis and resequencing of contactin associated protein-like 2 in autism spectrum disorders. *Am. J. Hum. Genet.* **82:** 165–173.
98. Ptashne, M. 2007. On the use of the word "epigenetic." *Curr. Biol.* **17:** R233–236.
99. Tan, E.K. 2007. The role of common genetic risk variants in Parkinson disease. *Clin. Genet.* **72:** 387–393.
100. Smith, S.S. *et al.* 1992. Genetic vulnerability to drug abuse. The D2 dopamine receptor Taq I B1 restriction fragment length polymorphism appears more frequently in polysubstance abusers. *Arch. Gen. Psychiatry* **49:** 723–727.
101. Persico, A.M. *et al.* 1996. D2 dopamine receptor gene TaqI A1 and B1 restriction fragment length polymorphisms: enhanced frequencies in psychostimulant-preferring polysubstance abusers. *Biol. Psychiatry* **40:** 776–784.
102. Goldberg, J. *et al.* 1993. A strategy for assembling samples of adult twin pairs in the United States. *Stat. Med.* **12:** 1693–1702.
103. Bucholz, K.K. *et al.* 1994. A new, semi-structured psychiatric interview for use in genetic linkage studies: A report on the reliability of the SSAGA. *J. Stud. Alcohol.* **55:** 149–158.
104. Itoh, K. *et al.* 2005. Association study between brain-derived neurotrophic factor gene polymorphisms and methamphetamine abusers in Japan. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **132:** 70–73.
105. Iwata, N. *et al.* 2004. No association is found between the candidate genes of t-PA/plasminogen system and Japanese methamphetamine-related disorder: A collaborative study by the Japanese Genetics Initiative for Drug Abuse. *Ann. N. Y. Acad. Sci.* **1025:** 34–38.
106. Ujiie, H. *et al.* 2004. [Genetic studies on substance dependence]. *Seishin Shinkeigaku Zasshi* **106:** 1598–1603.
107. Morita, Y. *et al.* 2005. A nonsynonymous polymorphism in the human fatty acid amide hydrolase gene did not associate with either methamphetamine dependence or schizophrenia. *Neurosci. Lett.* **376:** 182–187.
108. Nishiyama, T. *et al.* 2005. Haplotype association between GABAA receptor gamma2 subunit gene (GABRG2) and methamphetamine use disorder. *Pharmacogenomics J.* **5:** 89–95.
109. Hashimoto, T. *et al.* 2005. A functional glutathione S-transferase P1 gene polymorphism is associated with methamphetamine-induced psychosis in Japanese population. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **135:** 5–9.
110. Morita, Y. *et al.* 2005. The X-box binding protein 1 (XBP1) gene is not associated with methamphetamine dependence. *Neurosci. Lett.* **383:** 194–198.
111. Ikeda, M. *et al.* 2005. Positive association of AKT1 haplotype to Japanese methamphetamine use disorder. *Int. J. Neuropsychopharmacol.* 1–5.
112. Ide, S. *et al.* 2004. Gene polymorphisms of the mu opioid receptor in methamphetamine abusers. *Ann. N. Y. Acad. Sci.* **1025:** 316–324.
113. Harano, M. *et al.* 2004. A polymorphism of DRD2 gene and brain atrophy in methamphetamine psychosis. *Ann. N. Y. Acad. Sci.* **1025:** 307–315.

114. Inada, T. *et al.* 2004. No association found between the type 1 sigma receptor gene polymorphisms and methamphetamine abuse in the Japanese population: A collaborative study by the Japanese Genetics Initiative for Drug Abuse. *Ann. N. Y. Acad. Sci.* **1025:** 27–33.
115. Koizumi, H. *et al.* 2004. Association between the glutathione S-transferase M1 gene deletion and female methamphetamine abusers. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **126:** 43–45.
116. Ujike, H. *et al.* 2003. Nine- or fewer repeat alleles in VNTR polymorphism of the dopamine transporter gene is a strong risk factor for prolonged methamphetamine psychosis. *Pharmacogenomics J.* **3:** 242–247.
117. Carmelli, D. *et al.* 2002. Quantitative genetic modeling of regional brain volumes and cognitive performance in older male twins. *Biol. Psychol.* **61:** 139–155.
118. Lerman, C. *et al.* 2006. Role of functional genetic variation in the dopamine D2 receptor (DRD2) in response to bupropion and nicotine replacement therapy for tobacco dependence: Results of two randomized clinical trials. *Neuropsychopharmacology* **31:** 231–242.
119. David, S.P. *et al.* 2007. Pharmacogenetic clinical trial of sustained-release bupropion for smoking cessation. *Nicotine Tob. Res.* **9:** 821–833.
120. Rose, J.E. *et al.* 1998. Nicotine-mecamylamine treatment for smoking cessation: The role of precessation therapy. *Exp. Clin. Psychopharmacol.* **6:** 331–343.
121. Shifman, S. *et al.* 2008. A whole genome association study of neuroticism using DNA pooling. *Mol. Psychiatry* **13:** 302–312.
122. Barcellos, L.F. *et al.* 1997. Association mapping of disease loci, by use of a pooled DNA genomic screen. *Am. J. Hum. Genet.* **61:** 734–747.
123. Germer, S. *et al.* 2000. High-throughput SNP allele-frequency determination in pooled DNA samples by kinetic PCR. *Genome Res.* **10:** 258–266.
124. Hacia, J.G. *et al.* 1999. Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. *Nat. Genet.* **22:** 164–167.
125. Yang, Y. *et al.* 2003. Efficiency of single-nucleotide polymorphism haplotype estimation from pooled DNA. *Proc. Natl. Acad. Sci. USA* **100:** 7225–7230.
126. Moonesinghe, R. *et al.* 2008. Required sample size and nonreplicability thresholds for heterogeneous genetic associations. *Proc. Natl. Acad. Sci. USA* **105:** 617–622.
127. Hinds, D.A. *et al.* 2004. Application of pooled genotyping to scan candidate regions for association with HDL cholesterol levels. *Hum. Genomics* **1:** 421–434.
128. Risch, N. & J. Teng. 1998. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. *Genome Res.* **8:** 1273–1288.
129. Teng, J. & N. Risch. 1999. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. II. Individual genotyping. *Genome Res.* **9:** 234–241.
130. Hishimoto, A. *et al.* 2007. Neurexin 3 polymorphisms are associated with alcohol dependence and altered expression of specific isoforms. *Hum. Mol. Genet.* **16:** 2880–2891.
131. Ishiguro, H. *et al.* 2006. NrCAM in addiction vulnerability: Positional cloning, drug-regulation, haplotype-specific expression, and altered drug reward in knockout mice. *Neuropsychopharmacology* **31:** 572–584.
132. Dupont, W.D. & W.D. Plummer, Jr. 1990. Power and sample size calculations. A review and computer program. *Control Clin. Trials* **11:** 116–128.
133. Kruglyak, L. & E.S. Lander. 1995. High-resolution genetic mapping of complex traits. *Am. J. Hum. Genet.* **56:** 1212–1223.
134. Lander, E. & L. Kruglyak. 1995. Genetic dissection of complex traits: Guidelines for interpreting and reporting linkage results. *Nat. Genet.* **11:** 241–247.
135. Maller, J. *et al.* 2006. Common variation in three genes, including a noncoding variant in CFH, strongly influences risk of age-related macular degeneration. *Nat. Genet.* **38:** 1055–1059.
136. Li, M. *et al.* 2006. CFH haplotypes without the Y402H coding variant show strong association with susceptibility to age-related macular degeneration. *Nat. Genet.* **38:** 1049–1054.
137. Lin, D.Y. 2006. Evaluating statistical significance in two-stage genomewide association studies. *Am. J. Hum. Genet.* **78:** 505–509.
138. Gao, X. & J. Starmer. 2007. Human population structure detection via multilocus genotype clustering. *BMC Genet* **8:** 34.
139. Sabatti, C. *et al.* 2003. False discovery rate in linkage and association genome screens for complex disorders. *Genetics* **164:** 829–833.
140. Storey, J.D. & R. Tibshirani. 2003. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100:** 9440–9445.
141. Benjamini, Y. *et al.* 2001. Controlling the false discovery rate in behavior genetics research. *Behav. Brain Res.* **125:** 279–284.
142. Sawcer, S. *et al.* 1997. Empirical genomewide significance levels established by whole genome simulations. *Genet. Epidemiol.* **14:** 223–229.

143. Zhao, L.P. *et al.* 1999. On the assessment of statistical significance in disease-gene discovery. *Am. J. Hum. Genet.* **64:** 1739–1753.
144. Manolio, T.A. *et al.* 2007. New models of collaboration in genome-wide association studies: The Genetic Association Information Network. *Nat. Genet.* **39:** 1045–1051.
145. Wakefield, J. 2008. Reporting and interpretation in genome-wide association studies. *Int. J. Epidemiol.* **9:** 821–823.
146. Sonne, S.C. & K.T. Brady. 1999. Substance abuse and bipolar comorbidity. *Psychiatr. Clin. North Am.* **22:** 609–627, ix.
147. Shih, R.A. *et al.* 2004. A review of the evidence from family, twin and adoption studies for a genetic contribution to adult psychiatric disorders. *Int. Rev. Psychiatry* **16:** 260–283.
148. Antila, M. *et al.* 2007. Heritability of cognitive functions in families with bipolar disorder. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **144:** 802–808.
149. McGuffin, P. *et al.* 2003. The heritability of bipolar affective disorder and the genetic relationship to unipolar depression. *Arch. Gen. Psychiatry* **60:** 497–502.
150. Almeida, O.P. *et al.* 2002. Smoking as a risk factor for Alzheimer's disease: Contrasting evidence from a systematic review of case-control and cohort studies. *Addiction* **97:** 15–28.
151. Wolf, S.M. & J.P. Kahn. 2007. Genetic testing and the future of disability insurance: Ethics, law & policy. *J. Law Med. Ethics* **35**(2 Suppl): 6–32.
152. Kristman, V.L. & N. Kreiger. 2007. Information disclosure in population-based research involving genetics: A framework for the practice of ethics in epidemiology. *Ann. Epidemiol.* **18:** 335–341.
153. Lowrance, W.W. & F.S. Collins. 2007. Ethics. Identifiability in genomic research. *Science* **317:** 600–602.
154. Papassotiropoulos, A. *et al.* 2006. Common Kibra alleles are associated with human memory performance. *Science* **314:** 475–478.
155. Welcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447:** 661–678.
156. George, S.R. *et al.* 1993. Polymorphisms of the D4 dopamine receptor alleles in chronic alcoholism. *Biochem. Biophys. Res. Commun.* **196:** 107–114.
157. Muramatsu, T. *et al.* 1996. Association between alcoholism and the dopamine D4 receptor gene. *J. Med. Genet.* **33:** 113–115.
158. Li, T. *et al.* 2004. Association analysis of the DRD4 and COMT genes in methamphetamine abuse. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **129:** 120–124.
159. Welzl, H. & O. Stork. 2003. Cell adhesion molecules: Key players in memory consolidation? *News Physiol. Sci.* **18:** 147–150.
160. Benson, D.L. *et al.* 2000. Making memories stick: Cell-adhesion molecules in synaptic plasticity. *Trends Cell Biol.* **10:** 473–482.
161. Fredette, B.J. *et al.* 1996. Inhibition of motor axon growth by T-cadherin substrata. *Development* **122:** 3163–3171.
162. Chen, B.E. *et al.* 2006. The molecular diversity of Dscam is functionally required for neuronal wiring specificity in Drosophila. *Cell* **125:** 607–620.
163. Keene, A.C. *et al.* 2006. Drosophila dorsal paired medial neurons provide a general mechanism for memory consolidation. *Curr. Biol.* **16:** 1524–1530.
164. Kipmen-Korgun, D. *et al.* 2005. T-cadherin mediates low-density lipoprotein-initiated cell proliferation via the Ca(2+)-tyrosine kinase-Erk1/2 pathway. *J. Cardiovasc. Pharmacol.* **45:** 418–430.
165. Philippova, M. *et al.* 2005. RhoA and Rac mediate endothelial cell polarization and detachment induced by T-cadherin. *FASEB J.* **19:** 588–590.
166. Hug, C. *et al.* 2004. T-cadherin is a receptor for hexameric and high-molecular-weight forms of Acrp30/adiponectin. *Proc. Natl. Acad. Sci. USA* **101:** 10308–10313.
167. Ivanov, D.B. *et al.* 2001. Structure and functions of classical cadherins. *Biochemistry* (Mosc) **66:** 1174–1186.
168. Dean, C. *et al.* 2003. Neurexin mediates the assembly of presynaptic terminals. *Nat. Neurosci.* **6:** 708–716.
169. Yamakawa, K. *et al.* 1998. DSCAM: A novel member of the immunoglobulin superfamily maps in a Down syndrome region and is involved in the development of the nervous system. *Hum. Mol. Genet.* **7:** 227–237.
170. Kraus, D.M. *et al.* 2006. CSMD1 is a novel multiple domain complement-regulatory protein highly expressed in the central nervous system and epithelial tissues. *J. Immunol.* **176:** 4419–4430.
171. Takeuchi, T. *et al.* 2000. Expression of T-cadherin (CDH13, H-Cadherin) in human brain and its characteristics as a negative growth regulator of epidermal growth factor in neuroblastoma cells. *J. Neurochem.* **74:** 1489–1497.
172. Lein, E.S. *et al.* 2007. Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445:** 168–176.
173. Kim, Y.S. *et al.* 2003. GLIS3, a novel member of the GLIS subfamily of Kruppel-like zinc finger proteins with repressor and activation functions. *Nucleic Acids Res.* **31:** 5513–5525.
174. Barlow, G.M. *et al.* 2001. Down syndrome cell adhesion molecule is conserved in mouse and highly

expressed in the adult mouse brain. *Cytogenet. Cell Genet.* **94:** 155–162.
175. Wojtowicz, W.M. *et al.* 2007. A vast repertoire of Dscam binding specificities arises from modular interactions of variable Ig domains. *Cell* **130:** 1134–1145.
176. Gao, F.B. 2007. Molecular and cellular mechanisms of dendritic morphogenesis. *Curr. Opin. Neurobiol.* **17:** 525–532.
177. Hattori, D. *et al.* 2007. Dscam diversity is essential for neuronal wiring and self-recognition. *Nature* **449:** 223–227.
178. Koob, G.F. 2006. The neurobiology of addiction: A neuroadaptational view relevant for diagnosis. *Addiction* **101**(Suppl 1): 23–30.
179. Liu, F. *et al.* 2007. A genomewide screen for late-onset Alzheimer disease in a genetically isolated dutch population. *Am. J. Hum. Genet.* **81:** 17–31.
180. MacLaurin, S.A. *et al.* 2007. Hippocampal dendritic arbor growth in vitro: Regulation by Reelin-disabled-1 signaling. *Brain Res.* **1172:** 1–9.
181. Young-Pearse, T.L. *et al.* 2007. A critical function for beta-amyloid precursor protein in neuronal migration revealed by in utero RNA interference. *J. Neurosci.* **27:** 14459–14469.
182. Sheldon, M. *et al.* 1997. Scrambler and yotari disrupt the disabled gene and produce a reeler-like phenotype in mice. *Nature* **389:** 730–733.
183. Koop, E.A. *et al.* 2003. Receptor protein tyrosine phosphatase mu expression as a marker for endothelial cell heterogeneity; analysis of RPTPmu gene expression using LacZ knock-in mice. *Int. J. Dev. Biol.* **47:** 345–354.
184. Kee, H.J. *et al.* 2004. Expression of brain-specific angiogenesis inhibitor 3 (BAI3) in normal brain and implications for BAI3 in ischemia-induced brain angiogenesis and malignant glioma. *FEBS Lett.* **569:** 307–316.
185. Paul, S. & P.J. Lombroso. 2003. Receptor and nonreceptor protein tyrosine phosphatases in the nervous system. *Cell Mol. Life Sci.* **60:** 2465–2482.
186. Uetani, N. *et al.* 2000. Impaired learning with enhanced hippocampal long-term potentiation in PTPdelta-deficient mice. *EMBO J.* **19:** 2775–2785.
187. Uetani, N. *et al.* 2006. Mammalian motoneuron axon targeting requires receptor protein tyrosine phosphatases sigma and delta. *J. Neurosci.* **26:** 5872–5880.
188. Cai, H. *et al.* 2007. The sarcoglycan complex in Schwann cells and its role in myelin stability. *Exp. Neurol.* **205:** 257–269.
189. Shiga, K. *et al.* 2006. Zeta-sarcoglycan is a functional homologue of gamma-sarcoglycan in the formation of the sarcoglycan complex. *Exp. Cell Res.* **312:** 2083–2092.
190. Lau, W.L. & S.B. Scholnick. 2003. Identification of two new members of the CSMD gene family small star, filled. *Genomics* **82:** 412–415.
191. Zheng, C. *et al.* 1996. CNS gene encoding astrotactin, which supports neuronal migration along glial fibers. *Science* **272:** 417–419.
192. Adams, N.C. *et al.* 2002. Mice that lack astrotactin have slowed neuronal migration. *Development* **129:** 965–972.
193. Price, M. *et al.* 2003. Seven cDNAs enriched following hippocampal lesion: Possible roles in neuronal responses to injury. *Brain Res. Mol. Brain Res.* **117:** 58–67.
194. Cui, X.Y. *et al.* 2004. NB-3/Notch1 pathway via Deltex1 promotes neural progenitor cell differentiation into oligodendrocytes. *J. Biol. Chem.* **279:** 25858–25865.
195. Takeda, Y. *et al.* 2003. Impaired motor coordination in mice lacking neural recognition molecule NB-3 of the contactin/F3 subgroup. *J. Neurobiol.* **56:** 252–265.
196. Lee, S. *et al.* 2000. Expression and regulation of a gene encoding neural recognition molecule NB-3 of the contactin/F3 subgroup in mouse brain. *Gene* **245:** 253–266.
197. Hernandez-Montiel, H.L. *et al.* 2008. Semaphorins 3A, 3C, and 3F in mesencephalic dopaminergic axon pathfinding. *J. Comp. Neurol.* **506:** 387–397.
198. Ara, J. *et al.* 2004. Modulation of sciatic nerve expression of class 3 semaphorins by nerve injury. *Neurochem. Res.* **29:** 1153–1159.
199. Steup, A. *et al.* 2000. Sema3C and netrin-1 differentially affect axon growth in the hippocampal formation. *Mol. Cell Neurosci.* **15:** 141–155.
200. Park, C. *et al.* 2002. Deletion in Catna2, encoding alpha N-catenin, causes cerebellar and hippocampal lamination defects and impaired startle modulation. *Nat. Genet.* **31:** 279–284.
201. Uemura, M. & M. Takeichi. 2006. Alpha N-catenin deficiency causes defects in axon migration and nuclear organization in restricted regions of the mouse brain. *Dev. Dyn.* **235:** 2559–2266.
202. Caltagarone, J. *et al.* 2007. Focal adhesions regulate Abeta signaling and cell death in Alzheimer's disease. *Biochim. Biophys. Acta* **1772:** 438–445.
203. Fein, G. & V. Di Sclafani. 2004. Cerebral reserve capacity: Implications for alcohol and drug abuse. *Alcohol* **32:** 63–67.
204. Borenstein, A.R. *et al.* 2005. Developmental and vascular risk factors for Alzheimer's disease. *Neurobiol. Aging* **26:** 325–334.
205. Dean, C. & T. Dresbach. 2006. Neuroligins and neurexins: Linking cell adhesion, synapse formation and cognitive function. *Trends Neurosci.* **29:** 21–29.

206. Ushkaryov, Y.A. *et al.* 1992. Neurexins: Synaptic cell surface proteins related to the alpha-latrotoxin receptor and laminin. *Science* **257:** 50–56.
207. Ushkaryov, Y.A. *et al.* 1994. Conserved domain structure of beta-neurexins. Unusual cleaved signal sequences in receptor-like neuronal cell-surface proteins. *J. Biol. Chem.* **269:** 11987–11992.
208. Ushkaryov, Y.A. & T.C. Sudhof. 1993. Neurexin III alpha: Extensive alternative splicing generates membrane-bound and soluble forms. *Proc. Natl. Acad. Sci. USA* **90:** 6410–6414.
209. Boucard, A.A. *et al.* 2005. A splice code for trans-synaptic cell adhesion mediated by binding of neuroligin 1 to alpha- and beta-neurexins. *Neuron* **48:** 229–236.
210. Graf, E.R. *et al.* 2006. Structure function and splice site analysis of the synaptogenic activity of the neurexin-1 beta LNS domain. *J. Neurosci.* **26:** 4256–4265.
211. Graf, E.R. *et al.* 2004. Neurexins induce differentiation of GABA and glutamate postsynaptic specializations via neuroligins. *Cell* **119:** 1013–1026.
212. Kattenstroth, G. *et al.* 2004. Postsynaptic N-methyl-D-aspartate receptor function requires alpha-neurexins. *Proc. Natl. Acad. Sci. USA* **101:** 2607–2612.
213. Missler, M. *et al.* 1998 The making of neurexins. *J. Neurochem.* **71:** 1339–1347.
214. Missler, M. *et al.* 2003. Alpha-neurexins couple Ca2+ channels to synaptic vesicle exocytosis. *Nature* **423:** 939–948.
215. Rowen, L. *et al.* 2002. Analysis of the human neurexin genes: Alternative splicing and the generation of protein diversity. *Genomics* **79:** 587–597.
216. Tabuchi, K. & T.C. Sudhof. 2002. Structure and evolution of neurexin genes: Insight into the mechanism of alternative splicing. *Genomics* **79:** 849–859.
217. Occhi, G. *et al.* 2002. Identification and characterization of heart-specific splicing of human neurexin 3 mRNA (NRXN3). *Biochem. Biophys. Res. Commun.* **298:** 151–155.
218. Ullrich, B. *et al.* 1995. Cartography of neurexins: More than 1000 isoforms generated by alternative splicing and expressed in distinct subsets of neurons. *Neuron* **14:** 497–507.
219. Lachman, H.M. *et al.* 2007. Genomewide suggestive linkage of opioid dependence to chromosome 14q. *Hum. Mol. Genet.* **16:** 1327–1334.
220. Nissim-Rafinia, M. & B. Kerem. 2002. Splicing regulation as a potential genetic modifier. *Trends Genet.* **18:** 123–127.
221. Nissim-Rafinia, M. & B. Kerem. 2005. The splicing machinery is a genetic modifier of disease severity. *Trends Genet.* **21:** 480–483.
222. Gorecki, D.C. *et al.* 1994. Dystroglycan: Brain localisation and chromosome mapping in the mouse. *Hum. Mol. Genet.* **3:** 1589–1597.
223. Nikonenko, A. *et al.* 2003. Tenascin-R-deficient mice show structural alterations of symmetric perisomatic synapses in the CA1 region of the hippocampus. *J. Comp. Neurol.* **456:** 338–349.
224. Sorra, K.E. & K.M. Harris. 2000. Overview on the structure, composition, function, development, and plasticity of hippocampal dendritic spines. *Hippocampus* **10:** 501–511.
225. Grumet, M. 1997. Nr-CAM: A cell adhesion molecule with ligand and receptor functions. *Cell Tissue Res.* **290:** 423–428.
226. Niculescu, A.B., 3rd & J.R. Kelsoe. 2001. Convergent functional genomics: Application to bipolar disorder. *Ann. Med.* **33:** 263–271.
227. Ngandu, T. *et al.* 2007. Education and dementia: What lies behind the association? *Neurology* **69:** 1442–1450.
228. Uhl, G.R. & R.W. Grow. 2004. The burden of complex genetics in brain disorders. *Arch. Gen. Psychiatry* **61:** 223–229.
229. Ashburner, M. *et al.* 2000. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25:** 25–29.
230. Shapiro, L. *et al.* 2007. Adhesion molecules in the nervous system: Structural insights into function and diversity. *Annu. Rev. Neurosci.* **30:** 451–474.
231. Yamada, S. & W.J. Nelson. 2007. Synapses: Sites of cell recognition, adhesion, and functional specification. *Annu. Rev. Biochem.* **76:** 267–294.
232. Thiery, J.P. 2003. Cell adhesion in development: A complex signaling network. *Curr. Opin. Genet. Dev.* **13:** 365–371.
233. Rodriguez-Fernandez, J.L. & A.L. Corbi. 2005. Adhesion molecules in human dendritic cells. *Curr. Opin. Investig. Drugs* **6:** 1103–1111.
234. Hirano, S. *et al.* 2003. The cadherin superfamily in neural development: Diversity, function and interaction with other molecules. *Front. Biosci.* **8:** d306–355.
235. Bamji, S.X. 2005. Cadherins: Actin with the cytoskeleton to form synapses. *Neuron* **47:** 175–178.
236. Salinas, P.C. & S.R. Price. 2005. Cadherins and catenins in synapse development. *Curr. Opin. Neurobiol.* **15:** 73–80.
237. Mulder, N. & R. Apweiler. 2007. InterPro and InterProScan: Tools for protein sequence classification and comparison. *Methods Mol. Biol.* **396:** 59–70.
238. Wheeler, D.L. *et al.* 2008. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **36**(Database issue): D13–21.

239. Ohashi, K. *et al.* 1994. Cloning of the cDNA for a novel receptor tyrosine kinase, Sky, predominantly expressed in brain. *Oncogene* **9:** 699–705.
240. Brady-Kalnay, S.M. *et al.* 1995. Receptor protein tyrosine phosphatase PTPmu associates with cadherins and catenins in vivo. *J. Cell. Biol.* **130:** 977–986.
241. Brady-Kalnay, S.M. & N.K. Tonks. 1995. Protein tyrosine phosphatases as adhesion receptors. *Curr. Opin. Cell. Biol.* **7:** 650–657.
242. McAndrew, P.E. *et al.* 1998. Identification and characterization of RPTP rho, a novel RPTP mu/kappa-like receptor protein tyrosine phosphatase whose expression is restricted to the central nervous system. *Brain Res. Mol. Brain Res.* **56:** 9–21.
243. Beltran, P.J. & J.L. Bixby. 2003. Receptor protein tyrosine phosphatases as mediators of cellular adhesion. *Front Biosci.* **8:** d87–99.
244. DeFelipe, J. *et al.* 1999. Estimation of the number of synapses in the cerebral cortex: Methodological considerations. *Cereb Cortex* **9:** 722–732.
245. Schuz, A. & G. Palm. 1989. Density of neurons and synapses in the cerebral cortex of the mouse. *J. Comp. Neurol.* **286:** 442–455.
246. Posthuma, D. *et al.* 2003. Genetic correlations between brain volumes and the WAIS-III dimensions of verbal comprehension, working memory, perceptual organization, and processing speed. *Twin Res.* **6:** 131–139.
247. Posthuma, D. *et al.* 2002. The association between brain volume and intelligence is of genetic origin. *Nat. Neurosci.* **5:** 83–84.
248. Mori, S. & J. Zhang. 2006. Principles of diffusion tensor imaging and its applications to basic neuroscience research. *Neuron* **51:** 527–539.
249. Pfefferbaum, A. *et al.* 2001. Genetic regulation of regional microstructure of the corpus callosum in late life. *Neuroreport* **12:** 1677–1681.
250. Sullivan, E.V. *et al.* 2001. Heritability of hippocampal size in elderly twin men: Equivalent influence from genes and environment. *Hippocampus* **11:** 754–762.
251. Regier, D.A. *et al.* 1990. Comorbidity of mental disorders with alcohol and other drug abuse. Results from the Epidemiologic Catchment Area (ECA) Study. *JAMA* **264:** 2511–2518.
252. Stern, Y. *et al.* 1994. Influence of education and occupation on the incidence of Alzheimer's disease. *JAMA* **271:** 1004–1010.
253. Stern, Y. *et al.* 1999. Rate of memory decline in AD is related to education and occupation: Cognitive reserve? *Neurology* **53:** 1942–1947.
254. Bergeman, C.S. *et al.* 1993. Genetic and environmental effects on openness to experience, agreeableness, and conscientiousness: an adoption/twin study. *J. Pers.* **61:** 159–179.
255. Terracciano, A. & P.T. Costa, Jr. 2004. Smoking and the Five-Factor Model of personality. *Addiction* **99:** 472–481.
256. Brooner, R.K. *et al.* 1993. Antisocial personality disorder among drug abusers. Relations to other personality diagnoses and the five-factor model of personality. *J. Nerv. Ment. Dis.* **181:** 313–319.
257. Weitzdoerfer, R. *et al.* 2004. Neuronal nitric oxide synthase knock-out mice show impaired cognitive performance. *Nitric Oxide* **10:** 130–140.
258. Itzhak, Y. *et al.* 1998. The role of neuronal nitric oxide synthase in cocaine-induced conditioned place preference. *Neuroreport* **9:** 2485–2488.
259. Kleppisch, T. *et al.* 2003. Hippocampal cGMP-dependent protein kinase I supports an age- and protein synthesis-dependent component of long-term potentiation but is not essential for spatial reference and contextual memory. *J. Neurosci.* **23:** 6005–6012.
260. Liang, G.B. *et al.* 2007. Synthesis and SAR studies of potent imidazopyridine anticoccidial agents. *Bioorg. Med. Chem. Lett.* **17:** 3558–3561.
261. Asensio, A.C. *et al.* 2006. Biochemical and immunochemical characterisation of human diadenosine triphosphatase provides evidence for its identification with the tumour suppressor Fhit protein. *Biochimie* **88:** 461–471.
262. Pintor, J. *et al.* 2000. Diadenosine polyphosphate receptors. From rat and guinea-pig brain to human nervous system. *Pharmacol. Ther.* **87:** 103–115.
263. Weiske, J. *et al.* 2007. The tumor suppressor Fhit acts as a repressor of beta-catenin transcriptional activity. *Proc. Natl. Acad. Sci. USA* **104:** 20344–20349.
264. Grewal, P.K. *et al.* 2005. Characterization of the LARGE family of putative glycosyltransferases associated with dystroglycanopathies. *Glycobiology* **15:** 912–923.
265. Longman, C. *et al.* 2003. Mutations in the human LARGE gene cause MDC1D, a novel form of congenital muscular dystrophy with severe mental retardation and abnormal glycosylation of alpha-dystroglycan. *Hum. Mol. Genet.* **12:** 2853–2861.
266. Grewal, P.K. & J.E. Hewitt. 2002. Mutation of Large, which encodes a putative glycosyltransferase, in an animal model of muscular dystrophy. *Biochim. Biophys. Acta* **1573:** 216–224.
267. Hyman, S.E. 2005. Addiction: A disease of learning and memory. *Am. J. Psychiatry* **162:** 1414–1422.
268. Kelley, A.E. 2004. Memory and addiction: Shared neural circuitry and molecular mechanisms. *Neuron* **44:** 161–179.

269. Nestler, E.J. 2001. Neurobiology. Total recall—the memory of addiction. *Science* **292:** 2266–2267.
270. Mori, F. *et al.* 2000. Developmental changes in expression of the three ryanodine receptor mRNAs in the mouse brain. *Neurosci. Lett.* **285:** 57–60.
271. Alkon, D.L. *et al.* 1998. Time domains of neuronal Ca2+ signaling and associative memory: steps through a calexcitin, ryanodine receptor, K+ channel cascade. *Trends Neurosci.* **21:** 529–537.
272. Ross, O.A. *et al.* 2007. Advancing stroke therapeutics through genetic understanding. *Curr. Drug Targets* **8:** 850–859.
273. Giovannoni, M.P. *et al.* 2007. Synthesis of pyrrolo[2,3-d]pyridazinones as potent, subtype selective PDE4 inhibitors. *J. Enzyme Inhib. Med. Chem.* **22:** 309–318.
274. Zhang, H.T. *et al.* 2005. Effects of the novel PDE4 inhibitors MEM1018 and MEM1091 on memory in the radial-arm maze and inhibitory avoidance tests in rats. *Psychopharmacology (Berl.)* **179:** 613–619.
275. Verploegen, S. *et al.* 2005. Characterization of the role of CaMKI-like kinase (CKLiK) in human granulocyte function. *Blood* **106:** 1076–1083.
276. Balschun, D. *et al.* 1999. Deletion of the ryanodine receptor type 3 (RyR3) impairs forms of synaptic plasticity and spatial learning. *EMBO J.* **18:** 5264–5273.
277. Kouzu, Y. *et al.* 2000. Mutant mice lacking ryanodine receptor type 3 exhibit deficits of contextual fear conditioning and activation of calcium/calmodulin-dependent protein kinase II in the hippocampus. *Brain Res. Mol. Brain Res.* **76:** 142–150.
278. Ostrovskaya, O. *et al.* 2007. Inhibition of ryanodine receptors by 4-(2-aminopropyl)-3,5-dichloro-N,N-dimethylaniline (FLA 365) in canine pulmonary arterial smooth muscle cells. *J. Pharmacol. Exp. Ther.* **323:** 381–390.
279. Wang, Q. *et al.* 2000. Subcellular and developmental expression of alternatively spliced forms of fibroblast growth factor 14. *Mech. Dev.* **90:** 283–287.
280. Olsen, S.K. *et al.* 2003. Fibroblast growth factor (FGF) homologous factors share structural but not functional homology with FGFs. *J. Biol. Chem.* **278:** 34226–34236.
281. Vadasz, C. *et al.* 2007. Mesencephalic dopamine neuron number and tyrosine hydroxylase content: Genetic control and candidate genes. *Neuroscience* **149:** 561–572.
282. Adriani, W. *et al.* 2006. Short-term effects of adolescent methylphenidate exposure on brain striatal gene expression and sexual/endocrine parameters in male rats. *Ann. N. Y. Acad. Sci.* **1074:** 52–73.
283. Wozniak, D.F. *et al.* 2007. Impaired spatial learning and defective theta burst induced LTP in mice lacking fibroblast growth factor 14. *Neurobiol. Dis.* **26:** 14–26.
284. Xiao, M. *et al.* 2007. Impaired hippocampal synaptic transmission and plasticity in mice lacking fibroblast growth factor 14. *Mol. Cell Neurosci.* **34:** 366–377.
285. Dalski, A. *et al.* 2005. Mutation analysis in the fibroblast growth factor 14 gene: Frameshift mutation and polymorphisms in patients with inherited ataxias. *Eur. J. Hum. Genet.* **13:** 118–120.
286. Zhang, H.T. *et al.* 2002. Antidepressant-like profile and reduced sensitivity to rolipram in mice deficient in the PDE4D phosphodiesterase enzyme. *Neuropsychopharmacology* **27:** 587–595.
287. Miller, C.A. & J.F. Marshall. 2005. Molecular substrates for retrieval and reconsolidation of cocaine-associated contextual memory. *Neuron* **47:** 873–884.
288. Vorel, S.R. *et al.* 2001. Relapse to cocaine-seeking after hippocampal theta burst stimulation. *Science* **292:** 1175–1178.
289. Volkow, N.D. & T.K. Li. 2005. Drugs and alcohol: Treating and preventing abuse, addiction and their medical consequences. *Pharmacol. Ther.* **108:** 3–17.
290. Liu, X. *et al.* 1998. Smaller volume of prefrontal lobe in polysubstance abusers: A magnetic resonance imaging study. *Neuropsychopharmacology* **18:** 243–252.
291. Bolla, K.I. *et al.* 2005. Neural substrates of faulty decision-making in abstinent marijuana users. *Neuroimage* **26:** 480–492.
292. Bolla, K. *et al.* 2004. Prefrontal cortical dysfunction in abstinent cocaine abusers. *J. Neuropsychiatry Clin. Neurosci.* **16:** 456–464.
293. Grant, S. *et al.* 2000. Drug abusers show impaired performance in a laboratory test of decision making. *Neuropsychologia* **38:** 1180–1187.
294. Tsuang, M.T. *et al.* 2001. The Harvard Twin Study of Substance Abuse: What we have learned. *Harv. Rev. Psychiatry* **9:** 267–279.
295. Mortimer, J.A. *et al.* 2003. Head circumference, education and risk of dementia: Findings from the Nun Study. *J. Clin. Exp. Neuropsychol.* **25:** 671–679.
296. Godwin-Austen, R.B. *et al.* 1982. Smoking and Parkinson's disease. *J. Neurol. Neurosurg. Psychiatry* **45:** 577–581.
297. Allam, M.F. *et al.* 2004. Smoking and Parkinson's disease: Systematic review of prospective studies. *Mov. Disord.* **19:** 614–621.
298. Gottschalk, P.C. *et al.* 1999. Current concepts in pharmacotherapy of substance abuse. *Curr. Psychiatry Rep.* **1:** 172–178.

299. Tang, Y. *et al.* 2001. Total regional and global number of synapses in the human brain neocortex. *Synapse* **41:** 258–273.
300. Reiman, E.M. *et al.* 2007. GAB2 alleles modify Alzheimer's risk in APOE epsilon4 carriers. *Neuron* **54:** 713–720.
301. Feil, S. *et al.* 2005. Distribution of cGMP-dependent protein kinase type I and its isoforms in the mouse brain and retina. *Neuroscience* **135:** 863–868.
302. Demyanenko, G.P. *et al.* 2005. Abnormal neocortical development in mice lacking cGMP-dependent protein kinase I. *Brain Res. Dev. Brain Res.* **160:** 1–8.